Qualitative Representations for Recognition

by

Keith John Thoresz

B.S. in Computer Science, Rensselaer Polytechnic Institute, 1995
M.S. in Computer Science, University of Wisconsin-Madison, 1999

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Master of Science in Computational Neuroscience

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

July 2002

Author .....
Department of Brain and Cognitive Sciences
July 30, 2002

Certified by .....
Pawan Sinha
Assistant Professor
Thesis Supervisor

Accepted by .....
Earl Miller
Co-Chairman, Department Graduate Committee

Qualitative Representations for Recognition

by

Keith John Thoresz

Submitted to the Department of Brain and Cognitive Sciences
on July 30, 2002, in partial fulfillment of the
requirements for the degree of
Master of Science in Computational Neuroscience

Abstract

This thesis describes a representation for objects and scenes that is stable against variations in image intensity caused by illumination changes and tolerant to image degradations such as sensor noise. The representation, called a ratio-template, uses low-resolution ordinal contrast relationships as its matching primitives. The choice of these primitives was inspired not only by considerations of computational simplicity and robustness, but also by current knowledge of the early stages of visual processing in the primate brain. The resulting representation is biologically plausible, although there is currently no evidence to suggest that the representation is actually used by the primate visual system. Constructed manually at first, the ratio-template can be learned automatically from a set of examples. Two applications—face detection and scene indexing—are described. The ratio-template achieves detection rates higher than 90% and can process a 320×280 pixel image in 2.6 seconds at multiple scales.

Thesis Supervisor: Pawan Sinha
Title: Assistant Professor

Acknowledgments

There are many people whose help and encouragement I would like to acknowledge. I would like to start with my family, who has supported me unconditionally during my many years in school. Without my parents and my sister, my entire graduate career would have been far more difficult. In particular, my sister, Monique, provided many flights back home while I was stuck in Wisconsin. Now, I look forward to supporting her in her new golf career.

Next come my friends. First, the inner circle: Dāv Clark, Jodi Davenport, Amy Pooler, and Richard Russell. More than any others, you helped me see the light at the end of the tunnel and encouraged me at every step. You have been tremendous friends and I am glad that you will be in Cambridge for the next several years. My other friends have provided wonderful and necessary distractions along the way. To name a few: Carl Chu, Ronan Flynn, Haipeng Zheng, Kevin Helmick, Mary Prendergast, Meredith Talusan, Jason Elfenbein, and Winston Chang.

I would also like to thank Amy for introducing me to my winter passion: snowboarding. (And for keeping me awake with long chats on the sleep-deprived drives back from the mountains.) Laura Gomberg and Ginger Hoffman both receive kudos for the incredible balsamic.

Special thanks goes to Kevin Helmick and the MIT SCUBA Club for helping me to vigorously pursue one of my favorite hobbies. Kevin and I have spent many hours catching lobsters and visiting shipwrecks, including two German U-Boats, from his 15-foot Zodiac. I will never forget the harbor seals ("like puppy dogs") at the Isles of Shoals, the flying fish and the U-352, and the meteors coming back from a night at the Charles Haight.

Pawan Sinha, my thesis advisor, has played a large role in both my academic and personal life. After returning to MIT from Madison, where the work for this thesis began, Pawan encouraged me to apply to the Brain and Cognitive Sciences Department so that the research could continue. At MIT, he pushed me to participate in conferences to promote the work and keep it on track. When I first moved to

Cambridge, Pawan allowed me to stay at his apartment while I searched for a place of my own. He has hosted many parties at his old bachelor pad at 100 Memorial Drive, taken us out to innumerable lunches, and shared in my first (and his last) skydiving experience. I must also thank Pamela Lipson, Pawan's better half, for employing me and letting me borrow her computer for way too long.

Antonio Torralba, who's genius everyone in the computer vision world must eventually reckon with, was an incredible source of support. Antonio made several important technical contributions to the thesis, in addition to sporting an unflaggingly positive attitude and providing encouragement and stories about his tortured days as a graduate student.

But this thesis simply would never have been written if it were not for Denise Heintze, whose business card reads *Academic Administrator*, but whose efforts go well beyond that description. Denise is the glue that holds the department together. More than any other individual, she prevented me from walking away from the whole mess on several occasions, and in general has always looked out for my best interests.

Thanks everyone. If you want to find me, just look for the bubbles.

The author was supported in part by a National Institutes of Health Training Grant (T32-GM07484).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis describes a computational object representation that is suited to the problem of detecting objects in images. The representation, which is called a *ratio-template*, distinguishes itself from other object models in two ways. First, its photometric primitives were inspired by our knowledge of biological vision, demonstrating that biology can influence computational algorithms rather than just serve them passively as a performance benchmark. Second, the ratio-template is stable against changes in illumination and certain kinds of image degradation, such as sensor noise. Before describing the representation, the following section will briefly motivate my interest in this work.

1.1 Motivation

Object detection is the process of discriminating one class of objects, such as cars or faces, from all other object classes. This process is also referred to as between-class discrimination. Identification, on the other hand, is the process of discriminating familiar members within a particular object class, and is referred to as within-class discrimination. Both processes are clearly important, but detection is a prerequisite for identification and therefore serves as a good starting point for investigating computational object representations.

How are biologically-inspired algorithms informative and important to the field of

machine vision? Detection and identification are two fundamental tasks that humans perform countless times each day with no perceivable effort. Yet the visual world is richly populated and each view of it contains potentially hundreds of objects, most of which we are able to distinguish. Furthermore, we can recognize a familiar object or a scene despite many changes in its appearance, including changes in illumination, viewing direction, and occlusion. Little is known about how the human visual system accomplishes these feats. For example, what are the internal representations and mechanisms that drive our visual system's capabilities? The answers to this question offer a potentially rich set of tools for machine vision algorithms. Indeed, pieces of this puzzle have already led to algorithms for visual analysis [45, 49, 51]. This thesis describes a novel approach within this category of algorithms.

The remainder of this chapter will outline a few key approaches to object representation by previous researchers, and discuss the strengths and weaknesses of those approaches as well as their relevance to this work. The next chapter will introduce the ratio-template representation.

1.2 Prior work on object models for classification

A new idea does not develop in a vacuum. It is influenced by ideas that have come before it. To put this work in context and help highlight its contribution, this section provides a brief discussion of some of the relevant work by other authors.

1.2.1 Wavelets

Wavelet transforms are a method for decomposing functions into frequency components, called wavelets, similar to how a Fourier transform represents signals as a superposition of sine and cosine waves. Unlike Fourier transforms, wavelets are aperiodic and are localized in both frequency and space [10]. This makes them particularly useful for representing data with sharp discontinuities, such as edges. Wavelets are also useful for representing data at different scales. This is due to the fact that all wavelets are derived from a prototype function called an *analyzing wavelet* or *mother*

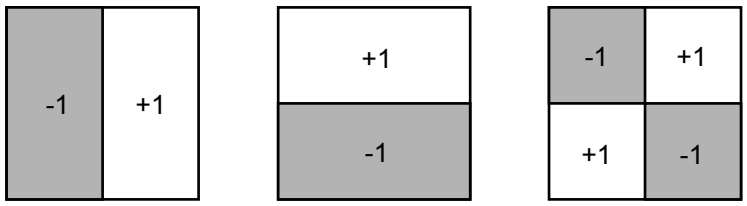


Figure 1-1: Oriented two-dimensional Harr wavelet filters. The gray regions represent areas where the wavelet function is negative, the white regions where the function is positive. The filters are applied to an image in search of features matching their description.

wavelet. Dilating or contracting the mother wavelet produces low- or high-frequency wavelets that can be used to analyze different types of information at different resolutions [27]. This flexibility lends itself easily to the analysis of images and their broad range of spatial information [46, 35, 41].

For an in-depth and formal discussion of wavelets, see Stollnitz *et al.* [46]. Briefly, the wavelet transform of an image is a summation of weighted terms. Some of the weights or coefficients will be very small relative to the others and may be eliminated without a significant loss of information, making wavelets a useful image compression technique [41]. Furthermore, operations on the transformed image need only involve the coefficients. The smaller the representation, the faster the computation.

Harr wavelets are the simplest type of wavelet and are used frequently for image analysis [39, 52, 41]. Stollnitz *et al.* [46] provide an excellent description of this wavelet basis. Based on square waves, Harr wavelets have certain desirable properties including compactness and symmetry. They are used as filters for analyzing image features at different orientations and scales. Figure 1-1 shows a schematic example of these filters.

Papageorgiou and Poggio [39] demonstrated an effective pedestrian detection system based on Harr wavelets. The system used an "overcomplete" set of wavelets, meaning that the wavelets were allowed to overlap so that the image would be more densely sampled. The features identified by the wavelets were fed to a support vector

machine (SVM)[1] learning algorithm. A total of 1,848 positive and 11,361 negative examples were used. The trained classifier model, containing approximate 25,000 features, was applied exhaustively to new images to detect the presence of pedestrians.

Wavelets were able to deal effectively with the spatial variability of the learning examples from this domain. The examples were normalized for scale but not for pose. Nor were the pedestrians segmented from their original backgrounds prior to learning. In addition to the overlapping wavelets, color information was used by applying the wavelet filters separately for each color channel and choosing the filter with the highest response. Initially the classifier model system included required 20 minutes to process a single image with a high classification rate. With performance optimizations, including feature selection—reducing the number of features to as little as 29 features—and the use of grayscale images, the system was capable of classifying roughly ten frames per second. However, classification performance degraded substantially as the speed increased.

Viola and Jones [52] used rectangular filters similar to Harr wavelets in a face detection application that is both fast and robust. Limited to vertical and horizontal orientations, the filters were used to model facial features. The filters—implemented as the absolute differences of average pixel values—were simpler than Harr wavelets and faster still, yet they provided a rich vocabulary for segmenting faces from their background. A learning algorithm chose those features that were of the greatest importance. The final detector included approximately 4300 features. It was able to operate at an extremely fast rate, scanning a 384×288 pixel image in 0.067 seconds and with a detection accuracy above 90%.

In summary, wavelets are an effective tool for representing image structure. The ratio-template scheme builds upon these ideas and leads to an encoding strategy that is simpler, more biologically plausible, and possibly more tolerant to image variations.

[1]SVMs are similar to neural networks. They learn iteratively by example and converge on a solution. SVMs use the strongest support vectors (i.e., training examples) to form a decision boundary. The classifying function is typically a linear combination of the support vectors.

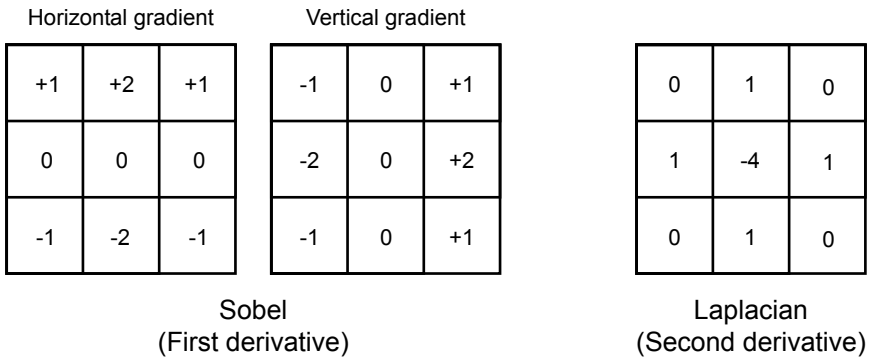


Figure 1-2: Examples of first and second derivative convolution operators. The Sobel operator approximates the first derivative. Sobel's anisotropic (directionally biased) kernels are applied separately for each edge direction. The isotropic Laplacian approximates the second derivative. These operators are convolved with an image and the output is thresholded to locate the edges.

1.2.2 Edge maps

Edges mark potentially important transitions in an image, including depth discontinuities, surface brightness (albedo) changes, and illumination boundaries. Edges are quickly and easily extracted from images using various edge detection algorithms, hence their popularity as an object representation. The simplest methods take either the first or second derivative of an image and look for, respectively, the maximum values or zero-crossings. The derivatives are easily computed by convolution operators like those shown in figure 1-2. The many variations of these operators differ in certain properties such as computational complexity and sensitivity to edges and noise.

Classification using edge maps typically proceeds as follows. A model edge map is extracted from an ideal instance of an object, such as that shown in figure 1-3. The map is compared to a set of edges extracted from a new image by measuring the distance between them. The distance is thresholded and the new object is classified as belonging to the model object class or not.

The critical parameters in an edge map-based detection system are the edge extraction algorithm and the distance measurement technique. The weakest link ap-

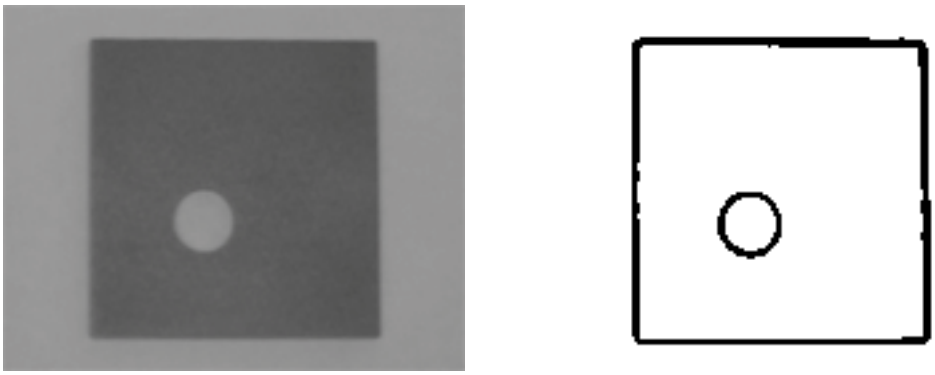


Figure 1-3: A model edge map (right) extracted from a grayscale image (left). The map can be used as a template to locate new instances resembling itself.

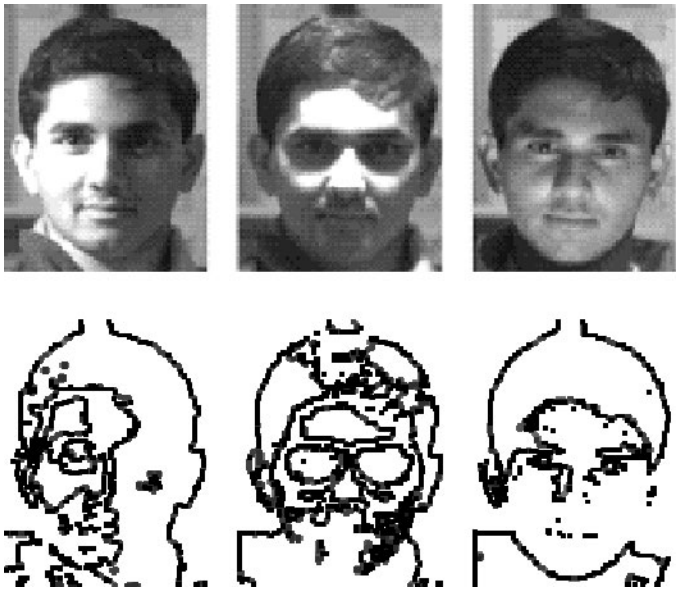


Figure 1-4: The same face under three different illumination conditions (above) and the resulting edge maps (below). This example demonstrates that edge maps can look very different depending on the distribution of image intensities, making recognition difficult.

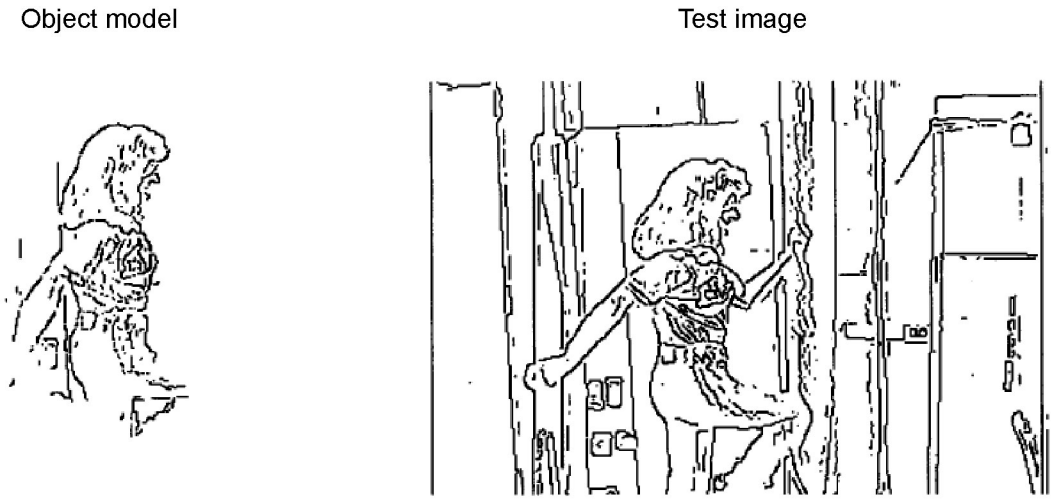


Figure 1-5: An edge map object model is used to locate a similar looking instance in the test image. From Huttenlocher *et al.* [32].

pears to be the edge detectors themselves. Although it is commonly believed that edges provide stable image features, in practice they are sensitive to changes in illumination, such as the direction and intensity of the light source. Changing lighting conditions can modify object boundaries significantly due to cast shadows and specularities. Consider figure 1-4. The edge maps for each face are dissimilar because the intensity boundaries have changed as the light source has shifted. The only stable feature across each of the faces is the external contour. A model based only on this feature would make recognition very difficult because much important information lies in the internal features of a face. It would be preferable to have a model that is invariant to illumination changes.

There are many ways to measure the distance between edge maps [32, 28, 38, 26, 12]. Huttenlocher *et al.* [32] provided one example. The distance from the edges in the image to those in the model is expressed by the Hausdorff metric. Consider two sets of points M and I belonging to the edge map model and the edges extracted from an image, respectively. The Hausdorff algorithm first locates the point $m \in M$ that is furthest from every point in I. The distance is then defined from m to the nearest point $i \in I$. The basic algorithm, modified to allow matching on partial

models to tolerate occlusion, was tested by matching object models to novel images containing instances of those objects (figure 1-5). Matching performance was good and outstripped simple correlation.

Some algorithms use deformable edge maps, also called active contours or snakes [34, 19]. These algorithms seek strong edges based on the image gradient and are guided by an energy minimization function or other objective function. Unfortunately, they suffer from the same edge segmentation problems as static edge maps.

A key assumption that motivates the use of edge maps is that the fine edge structure of an image is far more stable against illumination variations than the coarse, low-resolution distribution of intensities. The ratio-template approach turns this assumption around. It is based on the idea that the coarse photometric structure of an image provides a more stable pattern than the fine edge map.

1.2.3 Luminance templates

Luminance templates models [9, 6, 13, 7, 5, 14] are popular object models because they are conceptually intuitive. A luminance template is a set of one or more intensity maps (i.e., grayscale images) drawn from examples of the object class. To detect the presence of an object in an image, the template is correlated with the image and the correlation scores thresholded for matches.

Beymer [6] described a pose-invariant detection/recognition system that used templates comprising individual facial features (figure 1-6). The features were captured from a variety of people and winnowed by a clustering algorithm to yield templates with a degree of identity independence. The detector used a coarse-to-fine normalized correlation search strategy to locate the eyes and the nose. Detection and pose estimation took 10-15 minutes per individual, although the sluggishness of the system was due to the brute force pose estimator. (Each of the 62 people in the database was associated with 15 templates representing different viewpoints, as illustrated in figure 1-6. A total of 930 templates had to be correlated with each new image.) However, once detected, faces were recognized with high accuracy (above 94%).

The main problem with luminance templates is that correlation is highly sensitive

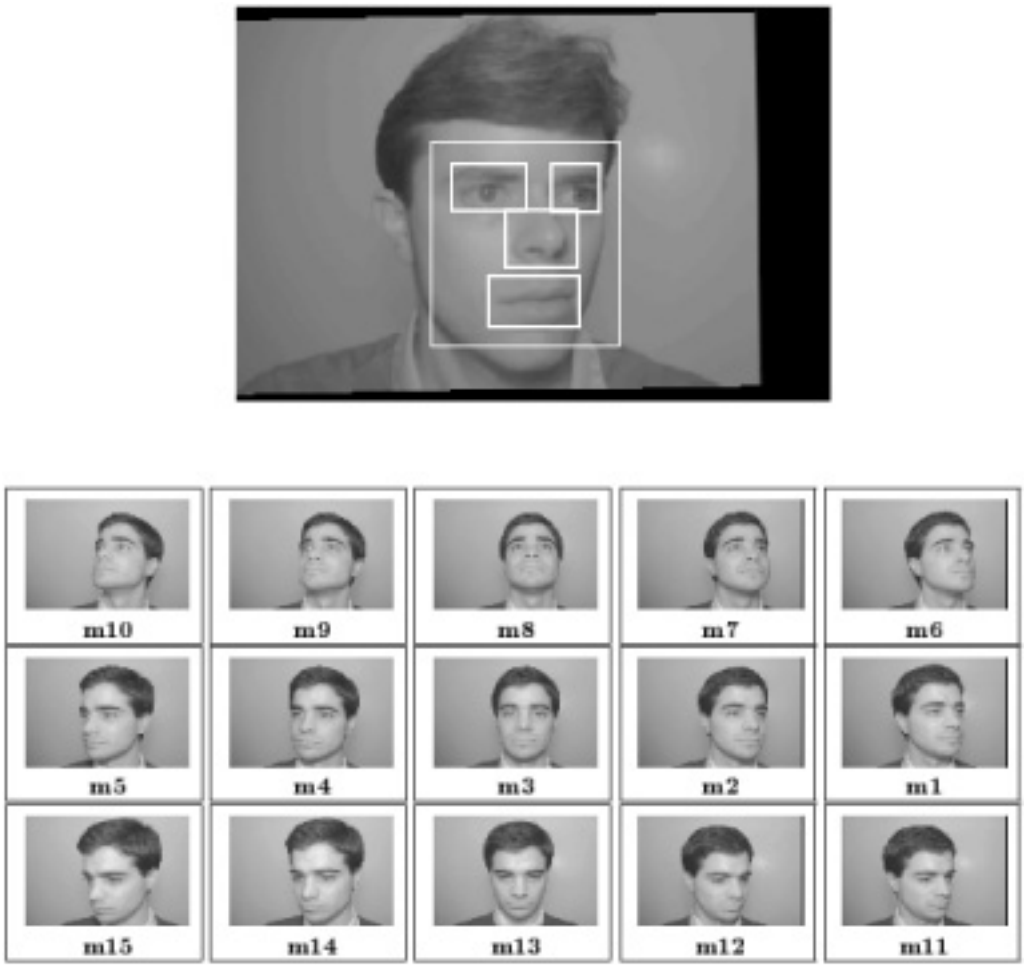


Figure 1-6: Templates used in Beymer's [6] face recognition system. (Top) White boxes highlight the features used as correlation templates. (Bottom) These templates were stored for fifteen views of each person in the database.

to changes in pixel intensity. To deal with this problem the templates and target images are often normalized prior to use [15, 3, 47, 6, 13]. One technique is to subtract the average pixel intensity of the template from each pixel and multiply by the template variance [3]. Other techniques use the image's gradient magnitude or Laplacian transform [13]. Beymer [6] and Brunelli [13] found that these techniques were only marginally helpful.

The difficulty in using luminance templates as effective representations is due to their intensity-sensitive correlation functions. To avoid this problem, the ratio-template scheme takes the opposite extreme: absolute intensities are ignored in favor of ordinal brightness relationships.

1.2.4 Color histograms

Object representations based on color histograms are widespread and used in applications such as image indexing and retrieval [23, 33, 31], recognition [17, 25, 21], and tracking [44]. Color histograms encode objects by the frequency of each pixel color in the image. This strategy encodes the global scene structure, making histograms sensitive to false positives due to contributions from image regions not belonging to the object of interest. Some implementations encode geometric information in order to overcome this limitation [17, 33, 31].

Swain and Ballard [48] developed a simple and effective recognition system based on color histograms that serves as the basis for many other such algorithms. First, a database was constructed in which each object was represented by both its histogram and the spatial areas of each color. Recognition was performed by intersecting the histograms of novel images with those of the objects in the database and choosing the best match. The matching algorithm was fast, requiring time proportional to the number of bins multiplied by the number of objects in the database. Its main disadvantage was its sensitivity to illumination.

Much work has been devoted to overcoming the illumination sensitivity of color histograms [23, 25, 21]. Funt and Finlayson [23] extended Swain's method by histogramming the ratios of pixel colors within local neighborhoods. Under the assump-

tion that illumination is locally constant, the ratios themselves are independent of the scene illumination. The method was successful on images where the scene illumination was varied, but performed slightly worse than Swain's method on images with fixed illumination.

Huang *et al.* [31] demonstrated a color histogram variant called a color *correlogram* that has been valuable for indexing and retrieving images from databases. Correlograms incorporate geometric information into the histogram representation by binning the spatial distance between pairs of pixels. For efficient computation, Huang used a subset of the correlogram, the *autocorrelogram*, which considers only similarly colored pairs of pixels. The correlogram (and autocorrelogram) combines both local and global image structure, making it less sensitive to local changes in object appearance and contributions from unrelated areas in the image. Similarity between images in the database and a query image was measured using a normalized absolute difference. A query on the database returned a series of images ranked by their similarity values. The results demonstrated that the autocorrelogram was tolerant to changes in object appearance due to different viewpoints. In most cases, the autocorrelogram outperformed the regular histogram.

Histogram-based techniques perform well in situations where different classes objects vary in coloration. However, they are of limited use for grayscale images when the different classes exhibit similar intensity frequencies and when the histograms for objects in a given class vary significantly under different viewing conditions. Under such conditions, a method for robustly encoding the photometric and geometric structure is needed. The ratio-template representation provides one such method.

1.3 Outline of the thesis

Chapter 2 details the ratio-template structure and its motivation from biology, and discusses two of its applications. Chapter 3 describes a learning algorithm for constructing the ratio-template automatically from a set of examples and presents results from simulations on faces. Finally, the thesis concludes with overall remarks on the

ratio-template design and performance, its strengths and weaknesses, and how future work can improve upon it.

Chapter 2

Qualitative representations of image structure

This chapter describes the ratio-template representation, originally developed for objects by Sinha [45] and later extended to scenes by Lipson [36]. Using primitives that were motivated by our knowledge of biological vision, the ratio-template is tolerant to variations in the appearance of the items it encodes. Specifically, it can accommodate changes in illumination, resolution and image noise. Changes in pose and orientation are not addressed by this representation.

Section 2.1 introduces the ratio-template's photometric primitives and describes their neurobiological origins. Section 2.2 details the ratio-template's structure and its construction in the context of a face detection task. Results of using the ratio-template to find faces in images are presented. Finally, the representation's extension to scenes and a relevant application—image database indexing—are discussed.

2.1 Low-resolution ordinal contrast relations

The previous chapter suggested that high-resolution image structure and absolute intensity values can lead to representations that are not robust to image transformations such as illumination changes and noise-based degradations. To address these problems, we propose the ratio-template model. The ratio-template representation

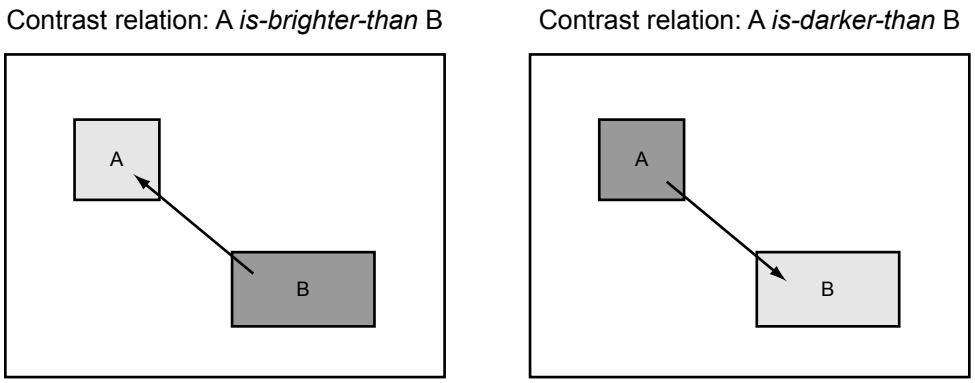


Figure 2-1: The definition of an ordinal contrast relation: one image region is either brighter than or darker than the other region. The exact brightness ratio is discarded and only the qualitative relationship is stored. In perceptual terms, qualitative relations are categorical relations.

is made robust to illumination changes and spatial degradation by its reliance on low-resolution ordinal contrast relations. The relations are part of a general class of ordinal operators, which act as qualitative filters by transforming a continuous space into a discrete, ordered space. School grades are one example of an ordinal operator where real valued grades $(0 - 100)$ are transformed into discrete ordered ones (A, B, C, D, F). The exact values underlying the letter grades are irrelevant. All that matters is the relative ordering of the discrete values. Ordinal operators are known for being robust to noise and outliers [18]. Suppose an application called for the school grades to be used in a correlation measurement. Using real valued grades, the values 80 and 89 would produce drastically different scores. Using the ordinal grade structure, any number in the range $[80, 89]$, once transformed into a letter grade, would have no effect on the score.

Bhat and Nayar [8] used ordinal operators in a window-based stereo image matching algorithm. Most stereo matching algorithms use some form of linear correlation, which performs poorly in the presence of outliers and intensity variation in and around corresponding pixels. Bhat and Nayar's algorithm converted the pixel values into intensity rankings and performed the matching on the rank values. The test images were degraded with different types of synthetic noise. In most tests, the ordinal oper-

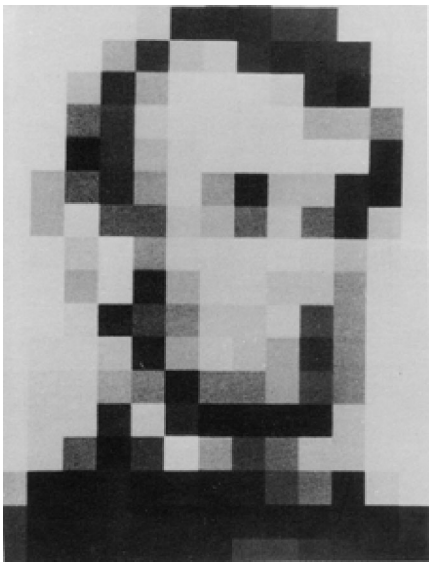


Figure 2-2: Low-resolution image of Abraham Lincoln demonstrates the importance of low-resolution features for recognition. From Harmon [30].

ator method outperformed every other method, including sum of squared differences, normalized cross correlation and a rank transform method by Zabih and Woodfill [54].

Based on ordinal operators, ordinal contrast relations are the photometric building blocks of the ratio-template. They provide qualitative measurements of the contrast magnitude and polarity (i.e., direction) between pairs of image regions. A binary ordinal value is used to encode the contrast magnitude as either above a given threshold (1) or not (0). Similarly, the contrast polarity between two image regions A and B is encoded as either *is-brighter-than* or *is-darker-than* (figure 2-1). The image regions involved in the contrast relations are purposely low-resolution, contributing to their robustness to noise and similar image degradations.

To summarize, ordinal contrast relations have several useful properties: they are conceptually simple, robust to image degradation and yield significant discriminatory power. What makes them even more compelling is evidence that they may be used by biological vision systems. The following sections present some of this evidence.

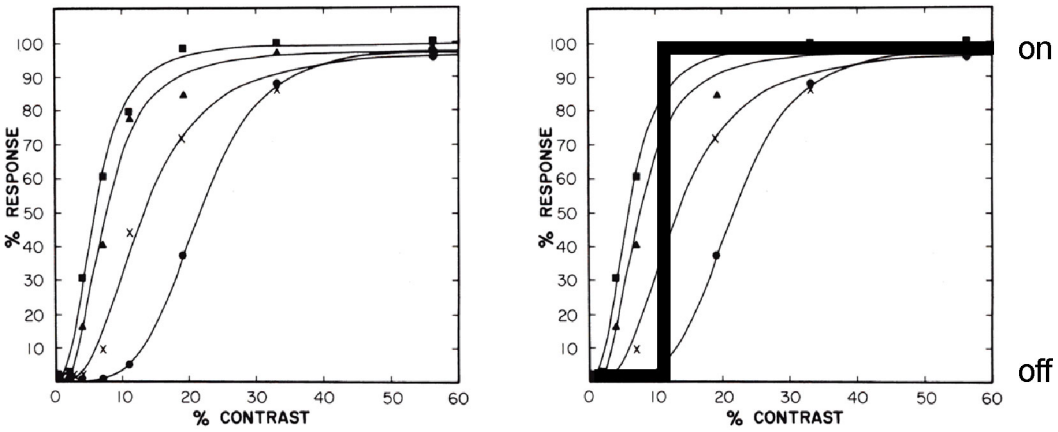


Figure 2-3: (Left) Cells in the primary visual cortex showing rapidly saturating contrast response functions. (Right) The contrast response function of such cells can be approximated by a step function. Beyond small contrast values, these cells provide information only about the polarity (direction) of the contrast and not its magnitude. In this way, they act as ordinal contrast filters. After Albrecht and Hamilton [2].

2.1.1 Biological support for low-resolution ordinal contrast relations

Biological vision systems may themselves use low-resolution ordinal contrast relations. Many cells in the primary visual cortex (V1) have large receptive fields [20] making them suitable for encoding low-resolution information. Anecdotal evidence of the biological importance of low-resolution features was shown by Harmon [30] in his now-famous picture of Abraham Lincoln (figure 2-2). A study by Torralba and Sinha [50] demonstrated that such features are sufficient, and in fact potent, for detection.

Both neurophysiological and perceptual studies support the idea of biological ordinal contrast filters. Albrecht and Hamilton [2] demonstrated that cells in V1 saturate rapidly in response to small values of contrast across the visual field (figure 2-3). The contrast response functions of such cells can be approximated by a step function: a cell either signals or not, depending on the size of the contrast magnitude. This two-state (off/on) function serves as a simple ordinal contrast filter. From the domain of visual perception, work by Cavanagh [16] and Watt [53] suggests that contrast polarity is important for basic object recognition. Cavanagh noted that the contrast resulting

Figure 2-4: Contrast polarity is important in recognizing objects when information about an object's contours is ambiguous or insufficient. After Cavanagh [16].

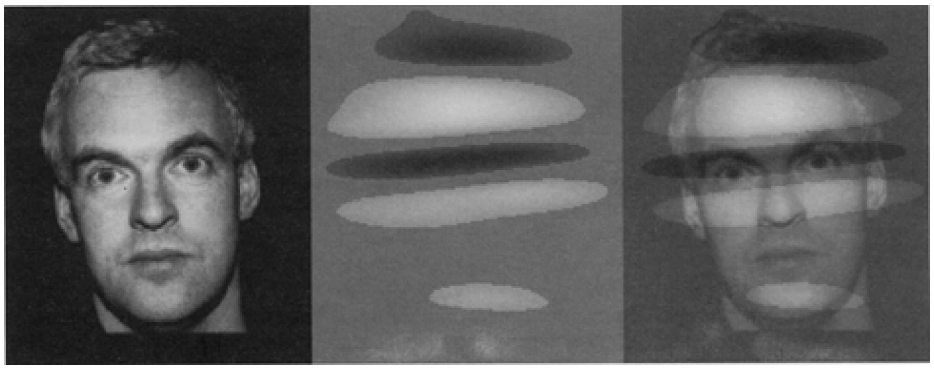


Figure 2-5: Watt [53] suggested that the barcode-like contrast pattern in faces facilitates the detection of faces in scenes.

from shadows is critical when information about object contours is ambiguous or otherwise insufficient (figure 2-4). Watt proposed that facial contrast patterns may be used like a barcode to quickly identify face-like objects (figure 2-5). Furthermore, many others have shown that photographic negatives of faces disturbs the recognition of individuals [11, 24, 40]. Although it appears that contrast polarity plays a defining role in our internal encoding of object structure, it is not clear that this structure is sufficient for recognition. The remainder of this chapter will demonstrate that it is.

2.1.2 Robustness versus discrimination

Although ordinal operators are inherently robust, they may not be appropriate for every application. They have tradeoffs like any other function. Ordinal contrast operators are used in the ratio-template representation because they are tolerant to noise and changes in illumination. But discarding the absolute magnitude of the contrast also discards information that could be useful for discrimination. For detection tasks this representation is sufficiently powerful. However, it is probably too weak to perform identification of individuals.

2.2 The ratio-template and two of its applications

2.2.1 Ratio-templates for face detection

With the ratio-template's photometric primitives defined, it is time to formulate the representation and discuss its first application: detecting faces. Faces are a compelling stimulus due to their ecological significance. They are also appropriate as a testbed for the ratio-template because they possess a stable characteristic structure. The object domain, shown in figure 2-6, is that of upright, frontal faces illuminated by a single light source positioned at different locations on a hemisphere around the face. A set of ordinal contrast relationships can be defined for these faces, but which ones should be selected? Because the representation is expected to be invariant to changes in illumination, relationships that are themselves invariant to illumination changes should be chosen. Figure 2-7 illustrates the selection process. Spatially coarse regions in each image are defined over salient facial features. Pairs of regions are inspected, and relations established between those pairs that exhibit consistent contrast polarity across all of the images. The pair-wise relations are combined into a 'hand-crafted' template that is schematically depicted in figure 2-8.

It should be noted that the idea of representing faces with relative brightness patterns is not new. Watt [53] proposed that the human face conforms to a bar-code-like brightness structure, as shown in figure 2-5. This is a special, less-refined case of

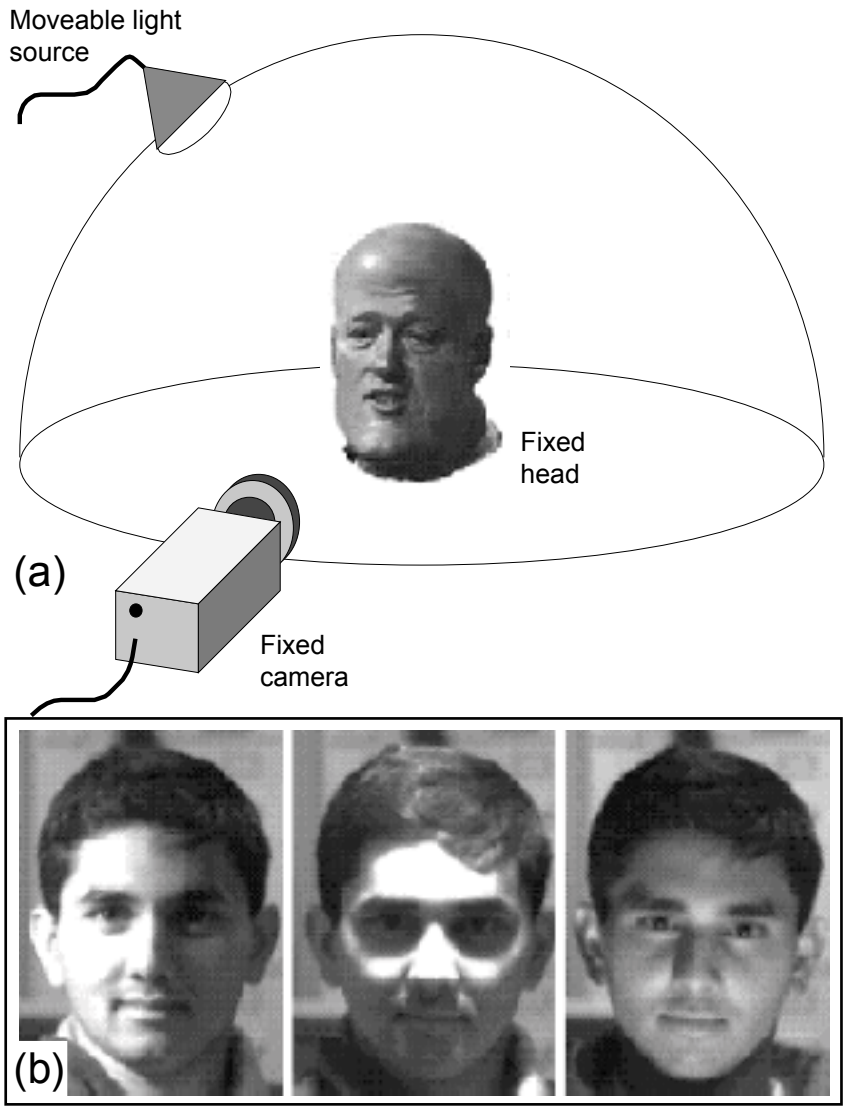


Figure 2-6: Object domain for the face ratio-template. (a) An upright, frontal face is illuminated by a single light source positioned on a hemisphere surrounding the head. (b) Three examples of the same face from this domain.

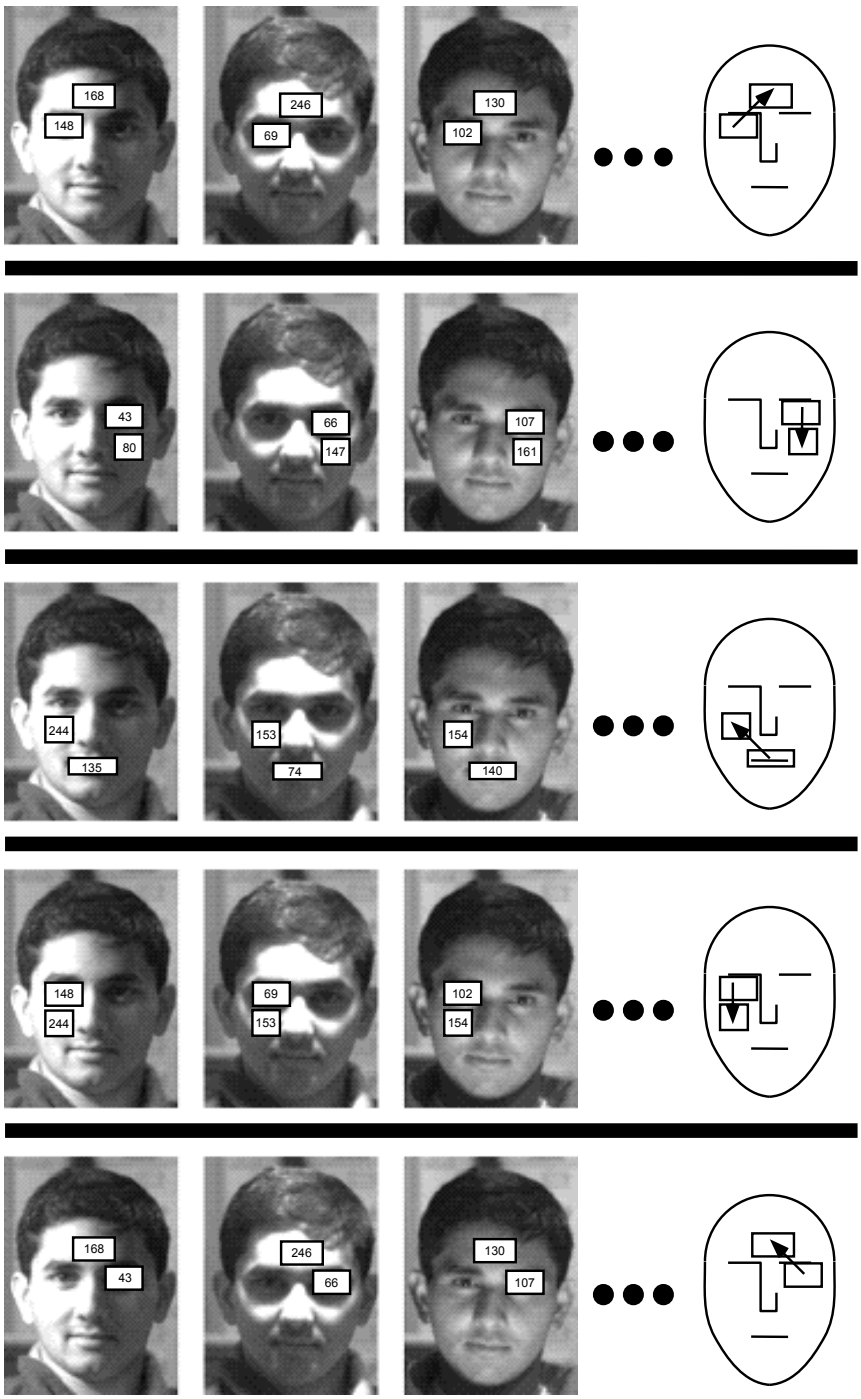


Figure 2-7: The derivation of five invariant contrast relationships from a face under three different illumination conditions. The numbers in the boxes are the average pixel intensities within those image regions. For each binary relation shown, the contrast direction remains unchanged across the different lighting conditions even though the intensity values (and the contrast magnitudes) change significantly. The ratio-template is constructed from a set of such relationships.

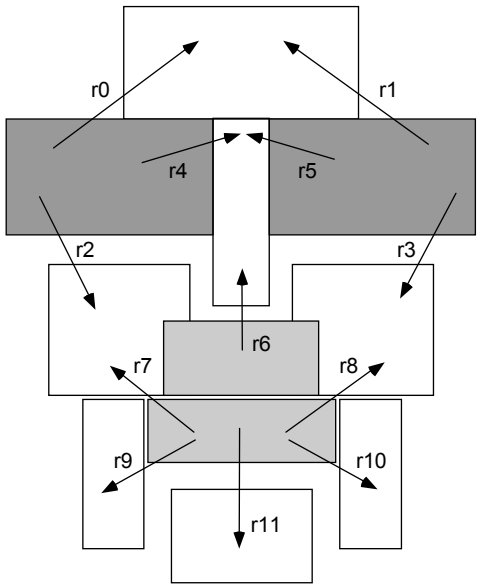


Figure 2-8: A hand-crafted template containing twelve ordinal contrast relations.

the ratio-template representation.

2.2.2 Match metric

Once constructed, the ratio-template can be used to detect faces in an image by exhaustively scanning the image with the template. At each position in the image, a set of ordinal contrast relations is extracted corresponding to the relations in the template. A face is detected when the image relations match enough of the template relations.

Let's look in more detail at how contrast relations are extracted from an image and how the image relations are matched against those in the template. The extraction process is illustrated in figure 2-9. Average pixel intensities are measured inside the image regions underlying the template, and a relation is created for each pair of intensities whose ratio is above a preset contrast threshold. (The contrast threshold is set at 20% to approximate the responses of the cells studied by Albrecht *et al.* [2].) The ordinal value of each image relation can then be compared to the value of its corresponding relation in the template.

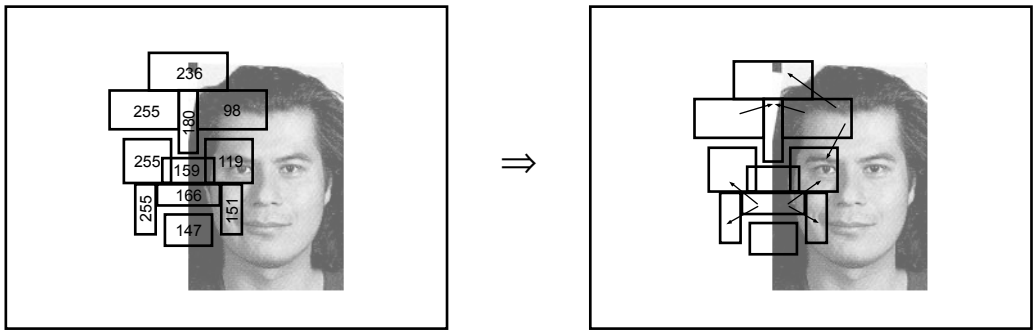
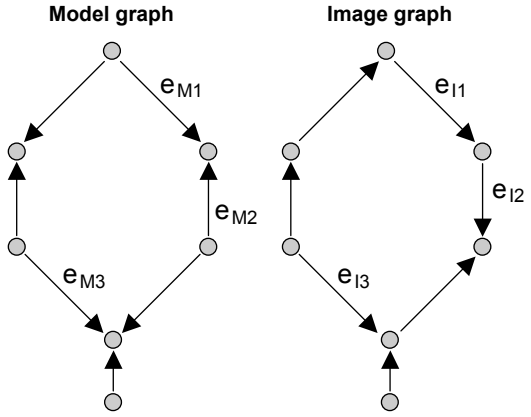


Figure 2-9: The ratio-template is used to scan an image for faces. At each position in the image, ordinal relations are extracted from the image and compared to the corresponding relations in the template. (Left) Average pixel intensities are measured inside the image regions underlying the template. (Right) Relations are created for pairs of intensities whose ratio is above a certain threshold. The relations are compared to those in the template, shown in figure 2-8, to determine whether a face is present.



$$\text{Match metric} \quad = \quad \sum_j f(e_{Mj}, e_{Ij})$$

$$\text{where} \quad f(e_{Mj}, e_{Ij}) = \begin{cases} 1.0 & e_{Mj} = e_{Ij} \\ -1.0 & e_{Mj} \neq e_{Ij} \end{cases}$$

Figure 2-10: Comparing the template to an image is analogous to matching two directed graphs. The nodes represent the image regions and the edges represent the ordinal contrast relationships. Corresponding edges in the graphs that have the same directions increment the match metric. Edges facing in opposite directions decrement the metric. For example, the graphs above produce a sum of $7 - 3 = 4$. If the match metric is close enough to its maximum value (i.e., when all edges match), a positive detection is made.

The matching of image relations to template relations can be generalized to a graph matching problem as shown in figure 2-10. The nodes in the graph represent image regions and the edges represent the ordinal contrast values. Whenever a pair of corresponding edges from the two graphs have the same edge direction, the match count is incremented. Corresponding edges with opposite directions decrement the count. A detection occurs if the match count is close enough to its maximum value. The number of matching relations required for detection depends on the desired power of the discrimination function. If a conservative template that yields the fewest false positives is desired, even at the expense of missing some actual faces, then the match threshold should be set at 100%. A more permissive template with a low match threshold will generate many false positives but miss few of the actual faces. This configuration would be suitable for a two-stage recognition algorithm whose second stage is more powerful but also more computationally expensive. The optimal match threshold is somewhere in between and must be determined experimentally. This detail is explored in the next chapter.

2.2.3 Results with faces

The ratio-template was tested on a set of frontal face images, shown in figure 2-11, that were scanned from various magazines. To measure the template's ability to avoid non-faces, two images with complex contrast textures containing no faces were included in the test set (figure 2-12). The receiver operating characteristics (ROC) curve in figure 2-13 provides an overview of the detector's performance. The ROC curve was generated by plotting the number of correct versus incorrect detections while varying the template's match threshold. The template performed well on this test set, detecting more than 95% of the faces while mislabelling as faces fewer than two in 10,000 images.

Figure 2-11: Test set of faces scanned from various magazines and normalized by the procedure shown in figure 3-6.

Figure 2-12: Two backgrounds with complex contrast textures used for performance benchmarking. The false positives found in these images were used in the performance curves for both the hand-crafted and learned ratio-templates.

2.2.4 Ratio-templates for scene indexing

Following work by Sinha [45], the ratio-template representation has been applied to other objects, including pedestrians [39] and cars [42]. Lipson [36] applied ratio-templates to natural and man-made scenes, such as waterfalls, snowy mountains and city skylines. Scenes differ from objects fundamentally in many ways. Consider figure 2-14. Objects are typically individual entities superimposed on a background, whereas scenes encompass the entire image, including the background and foreground. Objects tend to have a predictable or signature appearance. Scenes mostly refer to broad categories, whose instances can appear very different from one another in terms of color distribution, viewpoint, feature shapes, textures, etc. All of these differences pose a much greater challenge for the development of scene templates.

To help deal with these issues, Lipson introduced a range of features into the representation through the use of several types of ordinal operators, including

- Contrast operators, like those used in the face detector,

Detection performance of the hand-crafted ratio-template
on faces scanned from magazines

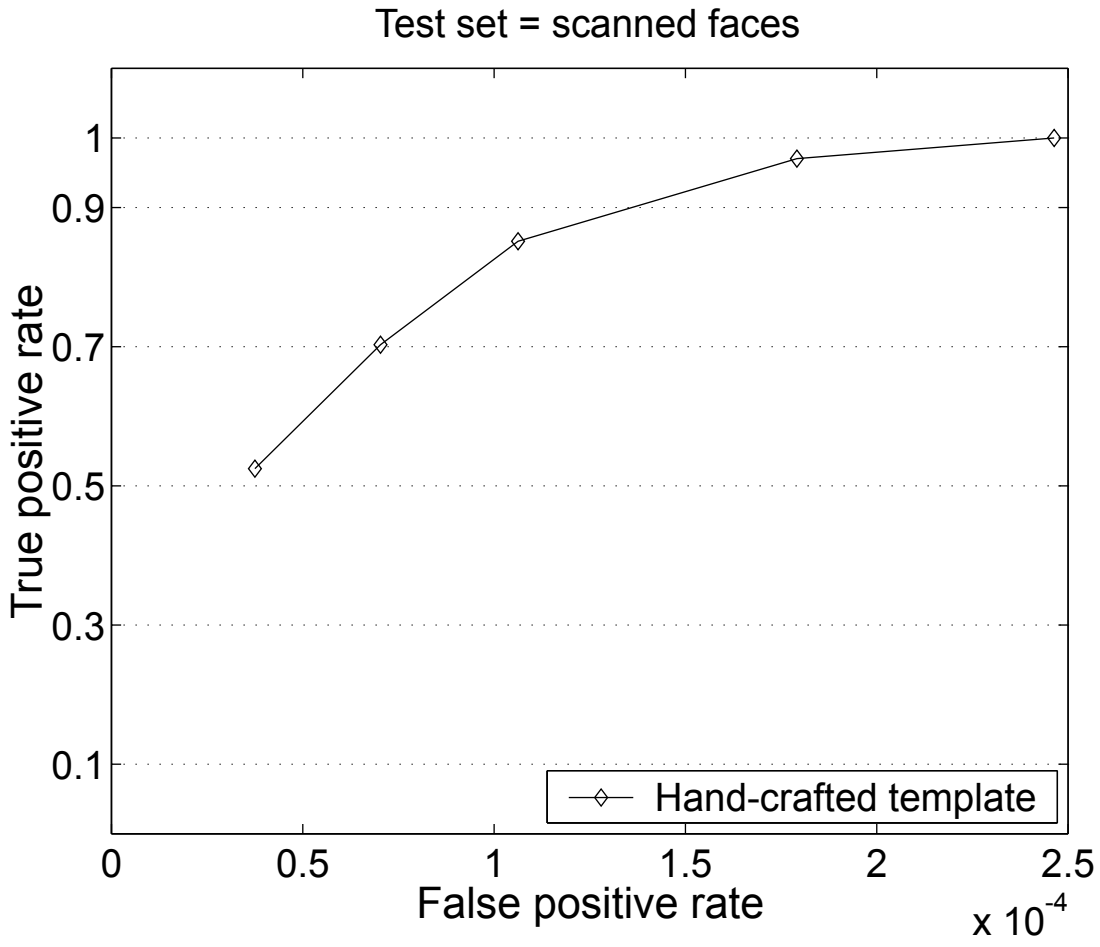


Figure 2-13: Receiver operating characteristics (ROC) curve showing the performances of the hand-crafted template on faces scanned from magazines. The curve was generated by plotting the number of correct versus incorrect detections while varying the template's match threshold. The template detected more than 95% of the faces while mislabelling as faces fewer than two in 10,000 images

Figure 2-14: Two examples each of an object (top) and a scene (bottom).

- Color operators, used to compare the red, green and blue color planes,
- Spatial operators, used to co-constrain the spatial locations of various features,
- Size operators, used to co-constrain the sizes of various features.

Increasing the number of features produces tradeoffs that must be managed. More features implies more dimensions along which scenes can be compared, thereby facilitating greater discrimination. But greater discrimination has potential costs. Each additional dimension provides an opportunity for the scene instances to become further separated. Excessive discrimination can lead to poor generalization, a problem known as overfitting [37]. Another problem with increasing the number of features is additional computational cost. This is an inexorable problem for all classification algorithms, and it can be managed only through careful feature selection and testing.

Finally, in order to tolerate variations in the scene geometry the scene templates were allowed to deform during the matching process. The template features were capable of moving independently within the constraints of the spatial operators.

Database indexing application

The scenes templates were applied to a database indexing application. A key requirement for a database is that its data be retrievable via user-specified queries. Naturally, the data must be assigned identities that can be specified in a query. The process of assigning these identities is called indexing. Indexing is often done manually, a tedious and error prone task at best. It is also common for the query interface to be specified in a symbolic language. Such an interface can be nonintuitive and cumbersome when the database stores a collection of images or other abstract data objects. It would be preferable to have a method that automatically indexes the objects and allows them to be queried using similar objects as examples.

Lipson's scene templates provide a mechanism for doing just that. The template representations serve as the image indices, which are generated once for all the images in the database. Queries are made by supplying an example image. The database engine generates a template for the query image and matches it against the stored templates, returning images in order of highest to lowest similarity. Lipson designed image models capable of discriminating snowy mountains, waterfalls, beaches, fields, and city skylines. A database containing approximately 700 images from many different categories was searched for images similar to the templates. The queries produced results as high as 80%-20% (true positives-false positives) for images of fields and as low as 33%-2% for waterfalls, with an average of 64%-6% across all the scene categories.

2.3 Summary

This chapter specified the structure of the ratio-template and provided neurobiological support for its underlying operators, the low-resolution ordinal contrast relations. Two application domains of the ratio-template, face detection and scene indexing, were discussed and results from tests of these applications were presented. Tests with faces demonstrated that the ratio-template is an effective classifier, achieving above a 95% detection rate with less than two false positives per 10,000 images.

The next chapter describes a learning system that constructs ratio-templates automatically from a set of training examples. The performance of the learned templates is demonstrated rigorously on a face detection task.

Chapter 3

Learning the ratio-template by example

This chapter describes a supervised learning system that constructs a ratio-template automatically from a set of examples. Given a set of images containing normalized instances of an object class, the learning system will encode the object class as a ratio-template. The template can then be used to detect instances of the class in novel images. The remainder of the chapter proceeds as follows. Section 3.1.1 discusses the training images. Section 3.1.2 describes how image features are generated. In section 3.1.3, the basic learning algorithm is outlined. Section 3.2 quantifies the learned template's performance on face detection tasks. Sections 3.2.1 to 3.2.4 present results demonstrating the template's tolerance to noise and illumination variations, and summarize tests of the algorithm's speed. Section 3.2.5 compares the performance of the learned templates with those of the hand-crafted template. Finally, section 3.2.6 discusses limitations of the learning system and suggests improvements.

3.1 Learning by example

To construct a data classifier by example requires an algorithm that can extract a *concept* from a set of training examples. A concept is a boundary (in some abstract space) that separates classes of objects from one another. For example, a face detec-

tion learning system extracts a face concept from the training examples by learning the boundary between the class of faces and everything else. A concept boundary can also separate multiple classes, an operation commonly referred to as clustering.

Training examples come in two main varieties: labeled and unlabeled. Labeled examples are marked individually as either belonging to the object class (positive) or not (negative). Learning with labeled examples is called *supervised learning* or learning with a teacher. The labels direct the learning algorithm toward the proper concept by showing it where the class boundary lies. All learning systems must use positive examples. Many use negative examples as well to help sharpen the boundary [6, 13, 7, 43]. Unlabeled examples are used in *unsupervised learning*, leaving the system to determine the boundary on its own by putting the examples into groups of similar likeness.

The ratio-template learning system is based on a supervised algorithm that takes as input a set of normalized example images. The images are searched systematically for invariant qualitative contrast relations that were described in chapter 2. The final set of relations constitute a ratio-template. Limited only by the training examples, the object domain of the learned template is similar to that of the hand-crafted template (figure 2-6). It is sensitive only to upright frontal faces, while being tolerant of spatial degradations such as image noise and variations in image intensity.

3.1.1 The training set

The ratio-template learning algorithm uses on the order of 100 normalized training examples; no negative examples are used. While many systems require thousands or tens of thousands of examples to learn [39, 12, 47], the ratio-template learning system can get by with a relatively small number because the contrast relations coarsely segment the feature space.

Two pairs of training and testing sets were derived from two image collections in the Harvard Face Database [29]. Already normalized, the faces in each collection were illuminated by a single light source that was placed at different positions relative to the camera's optical axis. Collection one contains 120 faces, which were divided into

Figure 3-1: Some of the faces from two collections of the Harvard Face Database [29]. (Top) Divided into Train Set 1 and Test Set 1, the faces in this collection were illuminated by a single light source that deviated in azimuth and elevation by at most 15 degrees. (Bottom) Divided into Train Set 2 and Test Set 2, and illuminated by a light source deviating by at most 30 degrees, the faces in this collection exhibit stronger shadows and other minor lighting artifacts.

96 training faces (Train Set 1) and 24 testing faces (Test Set 1). The light source in these images range in azimuth and elevation by at most 15 degrees. Collection two contains 180 faces, divided into 144 training faces (Train Set 2) and 36 testing faces (Test Set 2), with a range in azimuth and elevation of the light source by at most 30 degrees. Figure 3-1 shows some of the images from the two collections and demonstrates that the facial features in Train/Test Set 2 have been degraded due to the lighting conditions.

3.1.2 Extracting image features using the integral image

Like the hand-crafted template depicted in figure 2-8, the learned template is comprised of ordinal contrast relationships formed between pairs of rectangular image regions. To provide the learning system with a rich set of potential image features, regions of many different sizes must be sampled across the entire image. Fortunately, regions of arbitrary size are made readily available using the integral image representation described by Viola and Jones [52]. After a single-pass preprocessing operation, image regions can be extracted with just four memory accesses and three additive operations. A few extra multiplicative operations yields the average pixel value of a region, while squaring the pixels provides access to the variance of a region. As shown in figure 3-2, each pixel $ii(x, y)$ in the integral image contains the sum of the pixels in the region defined by its top-left corner $(0, 0)$ at the image origin and its bottom-right corner (x, y). More formally,

$$ii(x, y) = \sum_{x'=0}^{x} \sum_{y'=0}^{y} i(x', y'),$$

where $i(x', y')$ refers to the pixel values in the original image.

Gaining the speed benefits of the integral image requires that all the training images be transformed into their integral image representations prior to learning. This is done in a single pass per image using the recurrence relation

$$ii(x, y) = ii(x - 1, y) + ii(x, y - 1) - ii(x - 1, y - 1) + i(x, y),$$

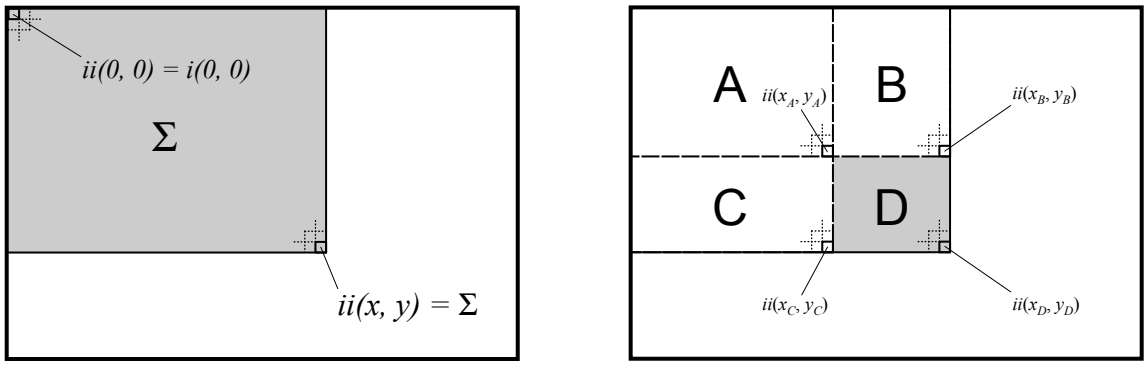


Figure 3-2: The integral image representation. (Left) The pixel $ii(x, y)$ in the integral image contains the sum $\sum$ of the pixels in the gray rectangular region bounded by coordinates $\{(0, 0), (x, y)\}$. (Right) An arbitrary region D is extracted by a trivial arithmetic operation on four pixel values.

with the boundary conditions

$$\begin{aligned} ii(x, -1) &= 0, \forall x \\ ii(-1, y) &= 0, \forall y. \end{aligned}$$

Figure 3-2 shows how the inverse transform

$$\sum_{x', y' \epsilon D} i(x', y') = ii(x_D, y_D) - ii(x_B, y_B) - ii(x_C, y_C) + ii(x_A, y_A)$$

is applied to extract an arbitrary region D from the integral image. It begins with the sum of the pixels $ii(x_D, y_D)$ in all four regions. The pixels in regions C and D are subtracted, causing the pixels in region A to be subtracted twice because both $ii(x_B, y_B)$ and $ii(x_C, y_C)$ contain them. Therefore, it is necessary to add back in the pixels from region A.

3.1.3 How the system learns and classifies

The goal of the learning system is to choose the best subset of contrast relations from the huge set of possible relations. The final set of relations should maximize

generalization to new instances of the learned object class and produce few false detections. The system takes as input a set of images containing normalized examples of the object class. Each image is transformed into its integral image representation. The system then exhaustively evaluates pairs of image regions to see whether each satisfies the *relation rule*. The relation rule states that

1. The contrast between a pair of regions must be greater than some threshold. That is, one region must be significantly brighter or darker than the other. A candidate relation is created when this is true.

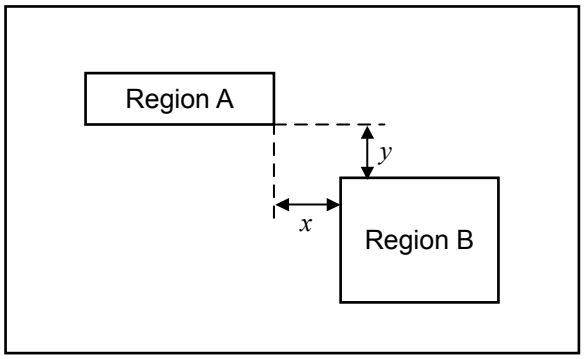
2. A candidate relation must be invariant across the training examples, meaning that the same relation in each image must have the same contrast polarity.

Candidate relations that satisfy the relation rule are included in the final template. Each relation is encoded as an ordinal value indicating the contrast relationship between its two regions (figure 2-1).

It is important to note that the ratio-template can only tolerate changes in illumination that it experiences from the training examples. Under normal conditions, the template would be unable to handle extreme lighting conditions. For example, illumination from below the face will reverse the contrast polarity of the facial features, making them incompatible with the template. In general, as illumination becomes more peripheral the representation will begin to break down.

Limiting the number of contrast relations

In the interest of execution speed it is necessary to limit the number of contrast relations prior to applying the relation rule. This is done by considering only a subset of all possible image regions made available by the integral image. Three techniques are used. First, the maximum and minimum dimensions of the image regions are constrained. For Train Set 1 and Train Set 2, the constraints were $\{width, height\} \in [2, 8]$. Second, the proximity of regions is constrained such that a relation may form between a pair only when they are sufficiently close (within two pixels) to one another.



Proximity constraint: $\{x < t, y < t\}$

Figure 3-3: A relation can form between two regions that are sufficiently close to one another. Proximity is determined by measuring the horizontal and vertical distances, x and y respectively, between the regions. The regions are sufficiently close if x and y are both less than threshold t. For our tests, $t = 2$.

$J(\mathbf{w}) > t$ $\quad$ $J(\mathbf{w}) \leq t$

Figure 3-4: Two examples of image region pairs evaluated by the Fisher linear discriminant function. The linear discriminant test is used to eliminate pairs of regions whose pixel distributions are weakly separated. The pixel values in each region are passed through a function $J(\mathbf{w})$ and thresholded. (Left) A pair of image regions whose pixel distributions are strongly separated. (Right) A pair of regions whose distributions strongly overlap.

Figure 3-3 shows how proximity is determined. Regions can and do overlap, yielding an overcomplete set of relations. It has been determined experimentally that an overcomplete set of relations is critical to a ratio-template's classifying ability.[1]

Finally, the remaining pairs of image regions must pass a linear discriminant test. This test eliminates pairs of image regions whose pixel distributions are not strongly separated. Linear discriminant analysis is often used in classification systems to separate two different distributions of data. The ratio-template learning system uses a variation of Fisher's [22] linear discriminant criterion function

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2},$$

where $\tilde{m}_1$ and $\tilde{m}_1$ are the means of the distributions and $\tilde{s}_1^2$ and $\tilde{s}_2^2$ are the variances. $J(\mathbf{w})$ projects high-dimensional data onto the line $\mathbf{w} \cdot \mathbf{x}$ and performs classification in a one-dimensional space.[2] The function maximizes the between-class separation and minimizes within-class scatter[3] over all linear projections $\mathbf{w}$ of the data [22]. We are interested in finding a minimal number of region pairs that have strongly separated pixel distributions, as shown in figure 3-4. By setting the threshold high enough, just below the performance drop-off point, it is possible to prune a significant number of relations while maintaining a high classification rate.

Classification

Classification of new stimuli is performed by sweeping the ratio-template across and down an image in a raster scan motion. At each location, the relations in the template are compared with those extracted from the underlying image. An instance of the object is found when enough relations match, as determined by the match threshold. The matching process was described in detail in section 2.2.2.

[1] Note that no attempt was made in this implementation to combine or eliminate regions that are potentially redundant due to a high degree of overlap. This optimization is addressed in chapter 4 as future work.

[2] The ratio-template system uses the un it projection $\mathbf{w} = \hat{\mathbf{i}}$ and the actual variances σ_1^2 and σ_2^2 rather than the scatter projections.

[3] Scatter refers to the sum of estimated variances

Input parameters				Ratio-templates		
Training set	**Size**	**Image dimensions**	**Fisher criterion**	**Relations**	**Dimensions**	**Legend**
Set 1	96	25×29	6.00	66	20×12	RT-6.0
Set 1	96	25×29	4.90	210	24×13	RT-4.9
Set 2	144	25×29	1.50	87	24×16	RT-1.5
Set 2	144	25×29	1.40	127	24×17	RT-1.4
Set 2	144	25×29	1.30	195	24×18	RT-1.3
Set 2	144	25×29	1.20	303	24×18	RT-1.2

Table 3.1: Ratio-templates produced by different Fisher criterion thresholds. The contrast threshold was set at 20% for all the templates. The training sets are shown in figure 3-1. The detection performances of the templates on new examples are compared in figures 3-5, 3-7, 3-8 and 3-9. (The templates can be cross-referenced by their legends.)

3.2 Results with faces

The following sections presents detection results for various ratio-templates trained on upright frontal faces. The discussion begins with a summary of the ratio-templates that were trained on Train Set 1 and Train Set 2, shown in figure 3-1. Performances on Test Set 1, Test Set 2 (section 3.1.1), the scanned faces and the non-face backgrounds (section 2.2.3, figures 2-11 and 2-12) are presented. Performance is demonstrated on the scanned faces in the presence of synthetic noise and across multiple spatial scales. Finally, the algorithm's speed is evaluated and a comparison is made to the hand-crafted template.

3.2.1 Preparing the ratio-templates

Several ratio-templates were trained on Train Set 1 and Train Set 2 while varying the Fisher criterion threshold. Because the Fisher criterion strongly influences the number and quality of relations in the template, it has a large effect on the template's detection performance. Although the contrast threshold also affects performance, it was fixed at 20%, which was the typical peak response level of the primary visual cortex cells studied by Albrecht *et al.* [2]. A summary of the templates is provided

in table 3.1.

3.2.2 Ceiling performance and generalization on two test sets

In the first round of tests, it was determined how well the templates could perform under ideal conditions by running them on their respective test sets, Test Set 1 and Test Set 2. These tests establish ceiling performance because the faces in these sets, though previously unseen, were drawn from the same population as the training images. Generalization to different lighting conditions was also measured by switching the test sets. The performance curves in figure 3-5 summarize the results. The templates perform well on their respective test sets, rapidly achieving 100% detection accuracy with few false positives. The switched test set results show that the ratio-templates generalized well to faces under different illumination, particularly when the illumination in the training set was more extreme than in the test set.

3.2.3 Tests on scanned faces

In order to test generalization to faces from different sources, the remainder of the tests (with one exception) described in this chapter were conducted on faces that were scanned from various magazines and normalized. Using manually labeled pupil locations, the normalization program uniformly scaled, uprighted and spatially registered the images (figure 3-6). Figure 3-7 compares the performance of three ratio templates on the scanned faces. All three templates reach greater than 90% detection accuracy while yielding fewer than one false positive per 1,000 images. This is one order of magnitude lower than the performances on Test Set 1 and Test Set 2.

Adding noise

One of the claims made of the ratio-template representation is that it is tolerant to image degradations such as noise. To test this hypothesis, one of the templates was applied to images degraded with 15% uniform noise. The results in figure 3-8 show that performance was affected only moderately, demonstrating that the template is

Detection performance of different ratio-templates

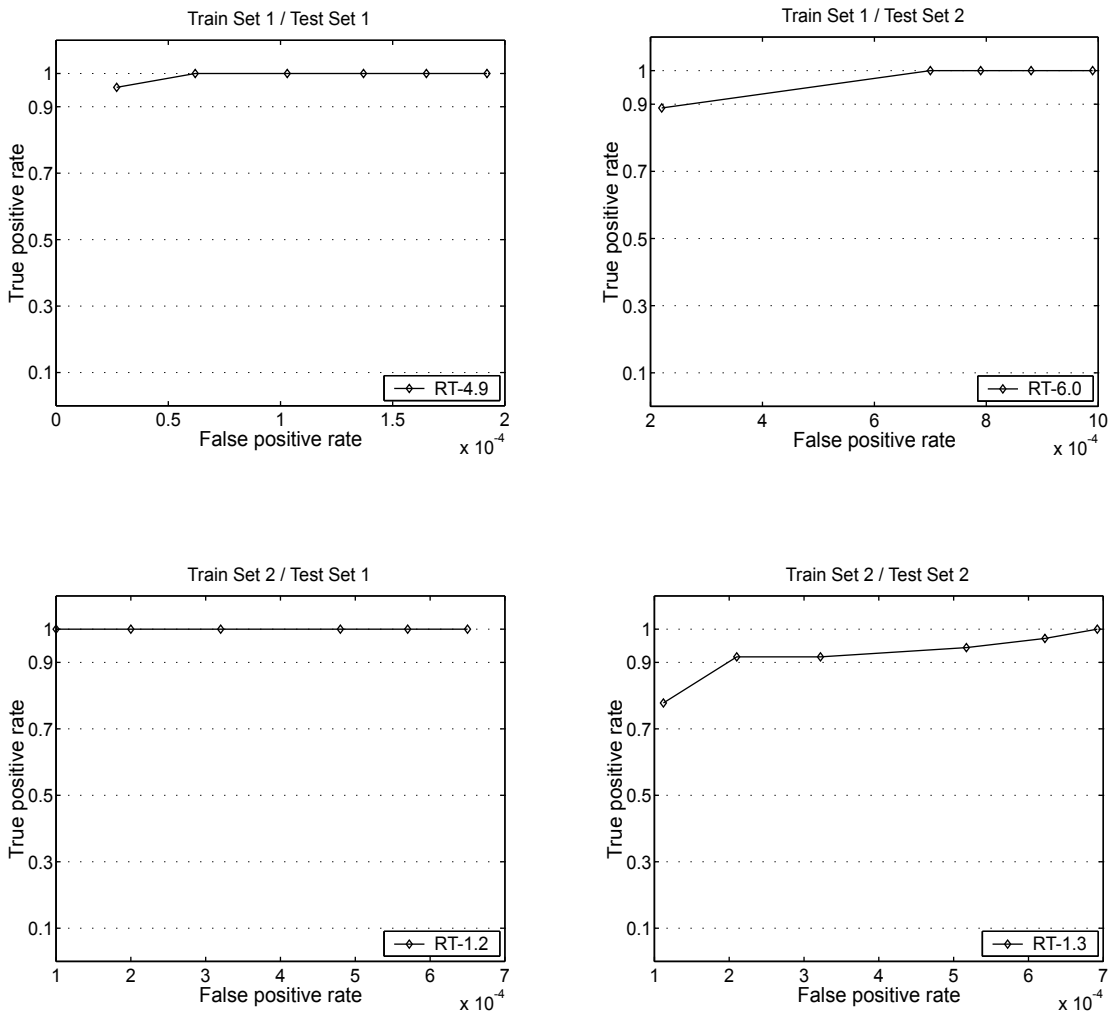


Figure 3-5: Receiver operating characteristics (ROC) curves comparing the detection performances of four ratio-templates summarized in table 3.1. The faces in Train/Test Set 1 and Train/Test Set 2 (figure 3-1) were differently illuminated. Set 1 has moderate non-frontal lighting. Set 2 has more extreme non-frontal lighting conditions, sometimes causing heavy shadows.

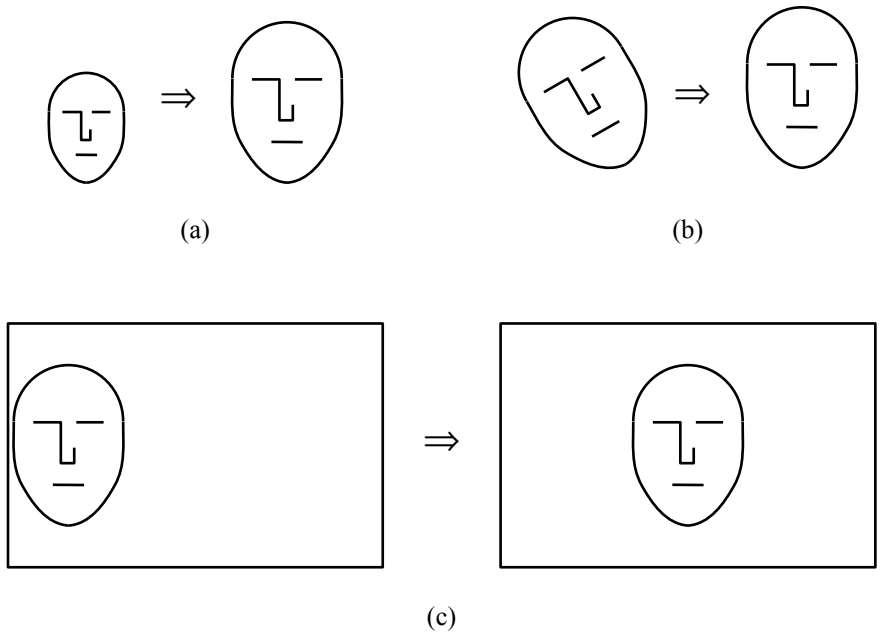


Figure 3-6: Normalization of faces using manually labeled pupil locations. (a) Faces are scaled to a predetermined size. (b) Uprighting tilted faces. (c) Registering the faces so that they are all located in the center of the image.

indeed robust to noise.

The raw data for the noisy and clean image tests are shown in table 3.2.

Detecting at multiple scales

The ratio-template is capable of detecting faces at multiple scales. This is achieved by scaling the template during the detection process. Typically, template matching schemes scale the image and not the template itself, resulting in large computational savings when convolving (i.e., matching) the template with the image [1]. However, when using the integral image representation, scaling the detector is faster than shrinking or expanding the image. This is because the integral image allows image windows of arbitrary size to be extracted in constant time. In other words, the time to analyze an image window is independent of its size.

The performance of the template was systematically tested on the scanned faces at multiple scales. The scanned faces and background images as a group were reproduced

Detection performance of different ratio-templates on scanned faces

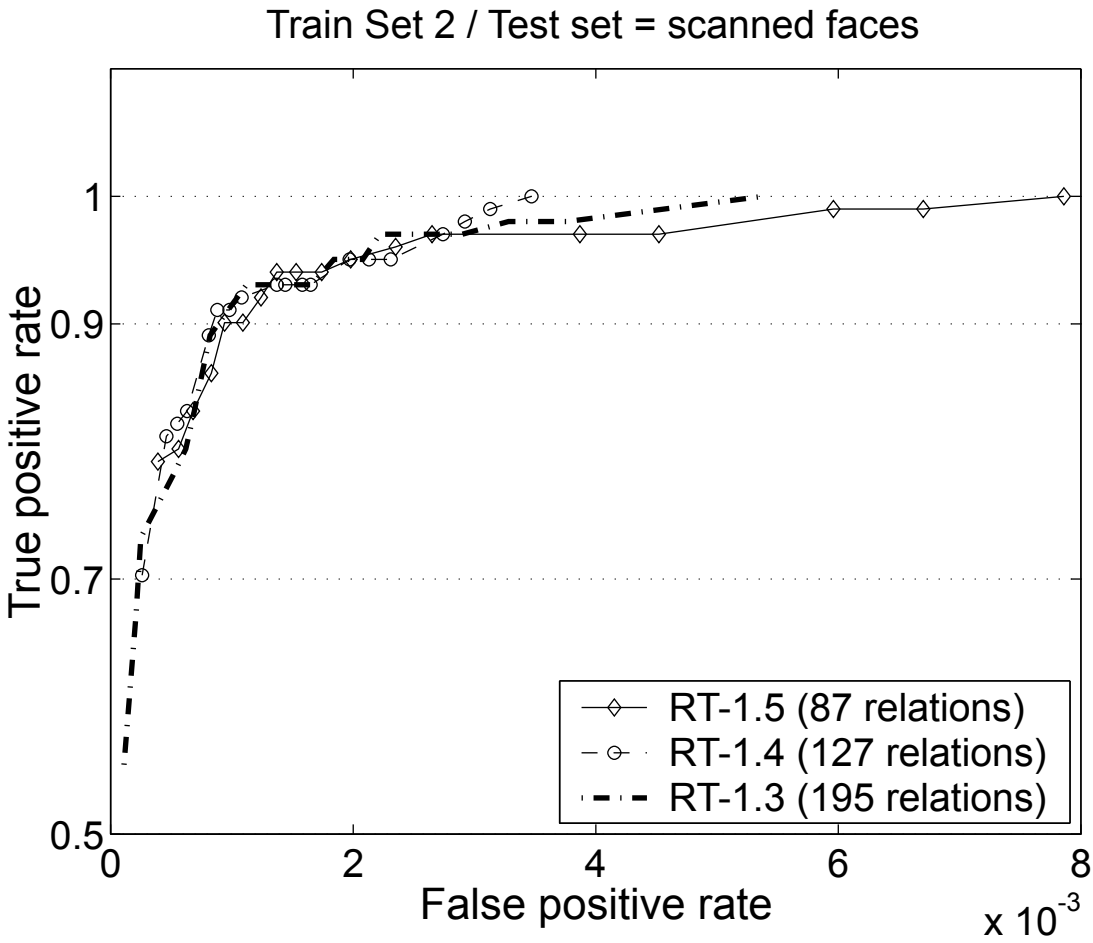


Figure 3-7: Receiver operating characteristics (ROC) curve comparing the performance of several different ratio-templates tested on faces that were scanned from magazines. The templates were trained on Train Set 2 using different Fisher criterion thresholds (see table 3.1). The best-performing template (RT-1.4) from this set has the optimal set of contrast relations.

Ratio-template detection performance on noisy images

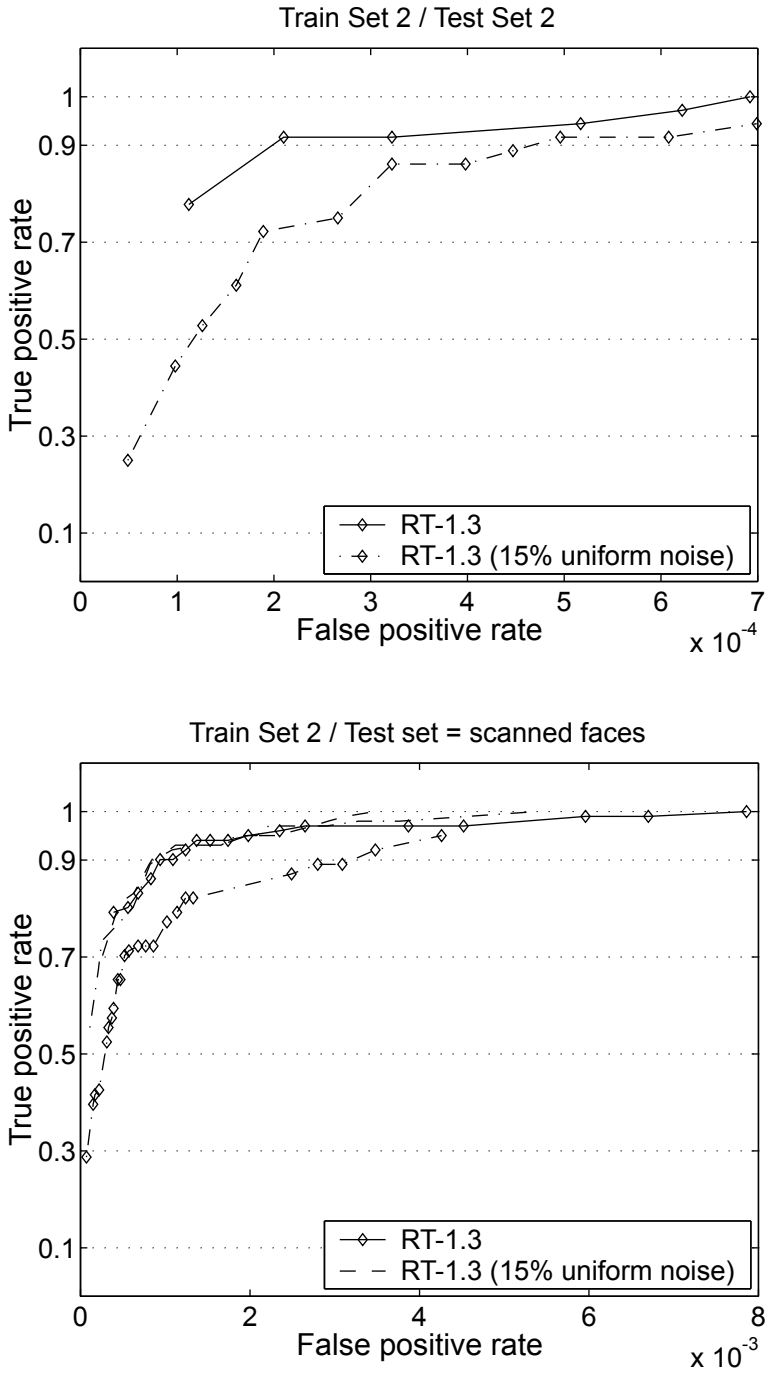


Figure 3-8: Receiver operating characteristics (ROC) curve showing the performance of two ratio-templates tested on the scanned faces degraded with 15% uniform noise. The noise causes a decrease in classification performance within acceptable levels, demonstrating that the ratio-template representation is tolerant to noise.

Ratio-template detection performance
(Train Set 2 / Test set = 101 scanned faces)

Match threshold (%)	**RT-1.3**		**RT-1.3 (noisy test data)**		**RT-1.4**		**RT-1.5**	
	TP	**FP**	**TP**	**FP**	**TP**	**FP**	**TP**	**FP**
100	56	16	29	10	71	38	80	56
99	68	30	40	22	82	66	81	81
98	74	46	42	25	83	79	84	98
97	79	74	43	31	84	91	87	119
96	81	89	53	45	90	116	91	136
95	84	99	56	47	92	127	91	157
94	90	117	58	53	92	140	93	178
93	92	139	60	56	93	155	95	197
92	94	160	66	63	94	197	95	197
91	94	182	66	67	24	207	95	220
90	94	211	71	75	94	227	95	250
89	94	238	72	82	96	237	96	285
88	96	263	73	97	96	283	97	339
87	96	298	73	111	96	306	98	382
86	98	318	73	124	96	331	98	557
85	98	369	78	146	98	393	98	651
84	98	415	80	164	99	419	98	651
84	99	469	83	178	100	449	100	859
83	99	538	83	191	101	498	100	965
82	101	770	88	357			101	1132

Table 3.2: Raw data for the detection performances of the various templates. To compute the false positive rates, divide the false positives by the number of windows scanned by each template: RT-1.30 (143,064), RT-1.40 (143,556), and RT-1.5 (144,048).

over a finite range of scales (60%, 80%, 120% and 140%), and the template was applied to each group at the same scale as the group. The template should have been able to detect all the faces at each scale, particularly since the scale was known a priori. The detection results in figure 3-9 show that the ratio-template is reasonably scale-invariant. Performance increases with scale as a result of the increasing pixel information in the image regions. More pixels means fewer outliers, better analysis of the pixel distribution by the Fisher linear discriminant, and more stable pixel averages from which the qualitative contrast ratios are formed.

3.2.4 System speed

The speed of the detection system is proportional to the number of contrast relations evaluated at each position in the image being scanned. This number is dependent on the similarity of the underlying image to the ratio-template as well as the order in which the relations are evaluated. For example, consider a ratio-template with 100 relations and the match threshold set at 95%. At each position in the image, as soon as five image relations failed to match the corresponding template relations the detector would move to the next position. Were those relations the first five inspected, then 95% of the relations would have been skipped and a large computational savings accrued. Conversely, were those relations the last five inspected, then no relations would have been skipped.

The system was used to scan a 340×240 pixel image across different match thresholds and different scales. A summary of the speeds was recorded in table 3.3. Without multiscale support and 90% detection accuracy the system scanned the image in 0.53 seconds. With multiscale support, the image was scanned in 2.6 seconds.

3.2.5 Comparison with the hand-crafted template

This section compares the performances of the hand-crafted template and the learned templates. In tests on the scanned faces, the hand-crafted template outperforms all of the learned templates. An example is shown in figure 3-10. The relative performances

Detection performance of a ratio-template at multiple scales

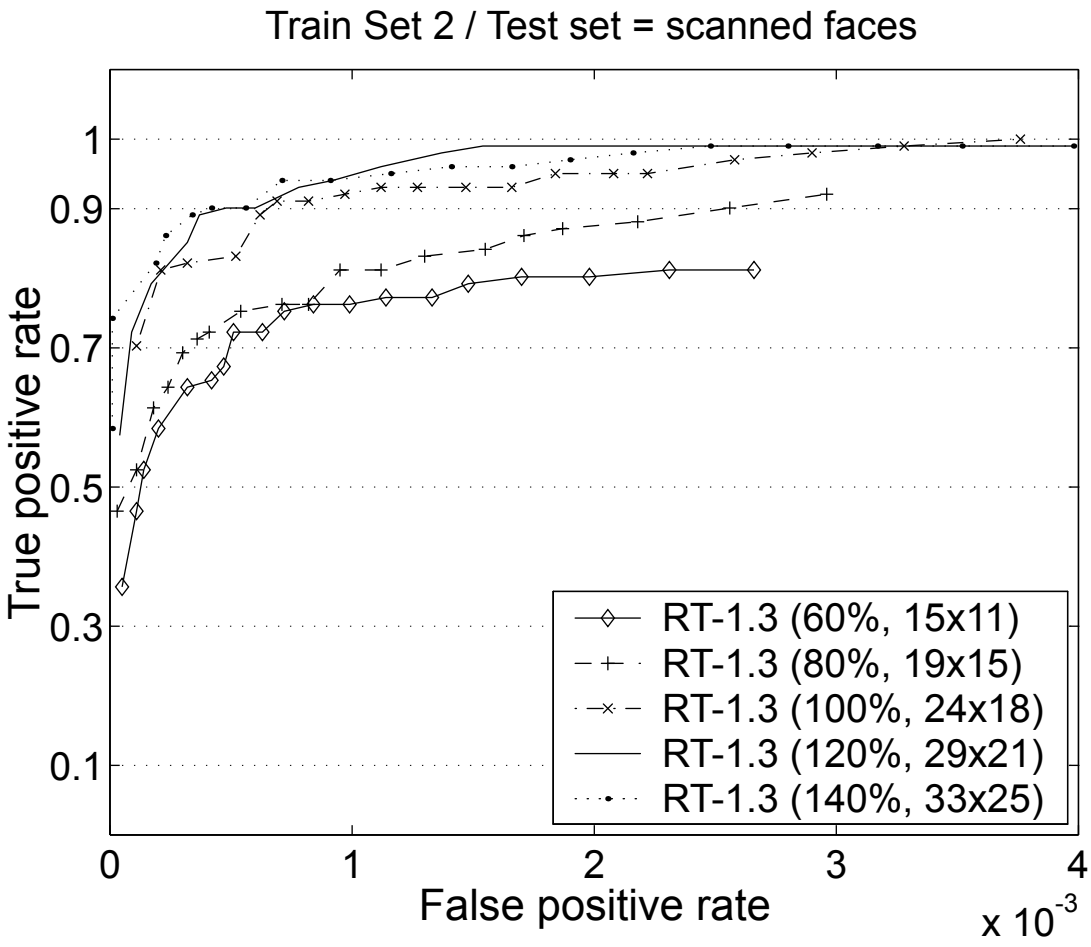


Figure 3-9: Receiver operating characteristics (ROC) curve showing the performance of a single ratio-template applied to the scanned faces at five different scales. The template performs better with increasing scale because because there is more pixel information available to produce more stable photometric measurements.

Comparison of the hand-crafted and learned templates

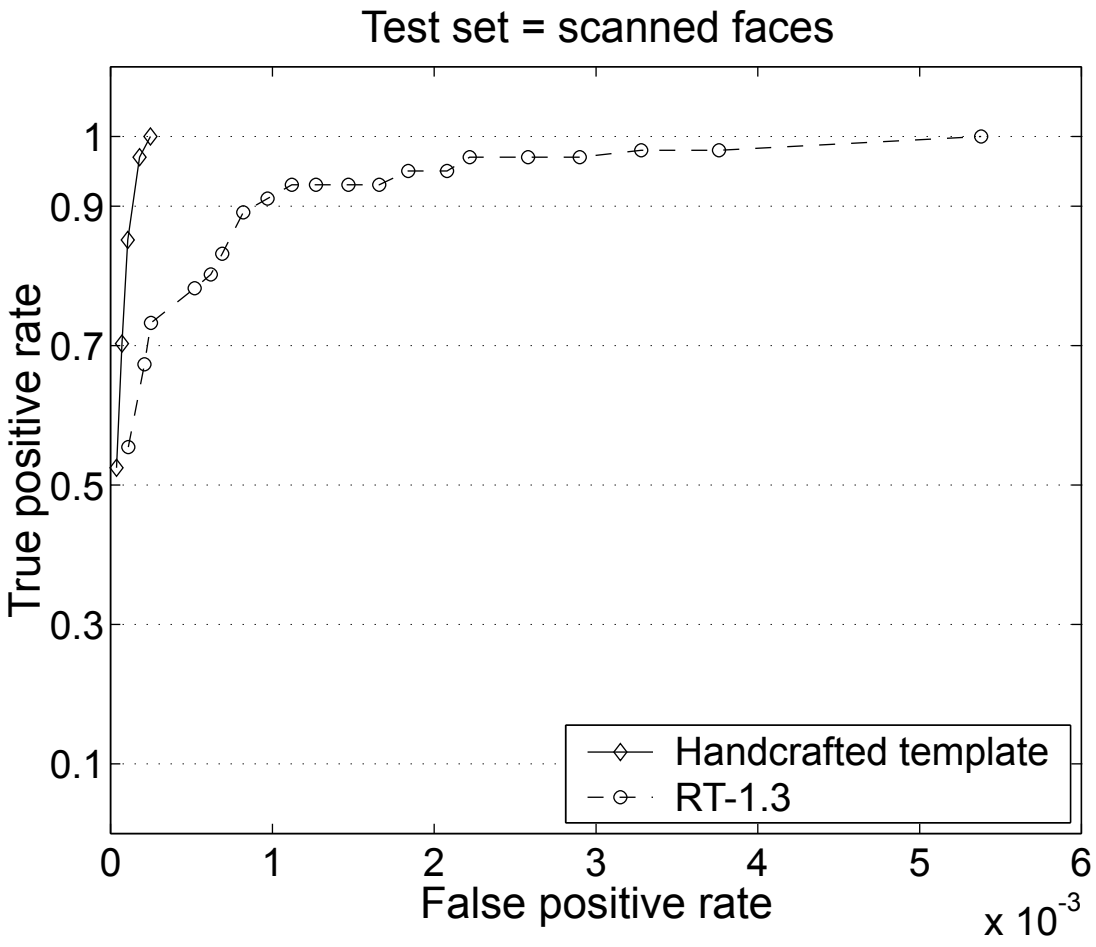


Figure 3-10: Receiver operating characteristics (ROC) curve comparing the performances of the hand-crafted and learned ratio-templates. Discussed in section 3.2.5, the relative performances suggest that the hand-crafted template is a better representation of the class of faces than the learned templates.

Ratio-template detection speed
(Train Set 2 / Test set = 101 scanned faces)

Match threshold (%)	RT-1.3 at one scale		RT-1.3 across multiple scales	
	Detection rate (%)	Speed (seconds)	Detection rate (%)	Speed (seconds)
100	55	0.046	54	0.19
99	67	0.098	65	0.45
98	73	0.14	71	0.68
97	78	0.19	75	0.91
96	80	0.25	80	1.2
95	83	0.29	82	1.4
94	89	0.33	82	1.6
93	91	0.38	85	1.9
92	93	0.44	86	2.1
91	93	0.48	87	2.4
90	93	0.53	89	2.6

Table 3.3: The speed of the detection system when applied to a 320×240 pixel image at different scales on a 1.7 GHz Pentium IV processor.

suggest that the hand-crafted template is a better representation of the class of faces than the learned templates. This is true despite the hand-crafted template containing significantly fewer relations.

The underlying problem with the learned templates might be an impoverished training set and/or a poor mechanism for selecting image features. Bear in mind that the hand-crafted template was designed by a creative human who, like all humans, has extensive experience with faces and was able to carefully consider the importance of each facial feature. On the other hand, the learned templates rely on the training examples and empirically determined parameters to select important features. Another possibility is that the features are being disproportionately emphasized by the normalization process, which focuses on the eyes (figure 3-6). In fact, some of the learned templates are known only to encompass the eye and forehead regions of the face. (By comparison, Rowley [43] uses a normalization technique that minimizes the sum of squared differences across a set of hand labeled features.)

One solution would be to use a different training set. Determining an optimal set

of parameters, including the contrast threshold, match threshold and Fisher criterion would require systematic exploration. It might also be worth extending the overcompleteness of the contrast relations by relaxing the image region constraints discussed in section 3.1.3.

3.2.6 Limitations and optimizations

There are two main limitations in the current implementation of the learned ratio-templates. First, classification performance is low compared to both the hand-crafted template and state-of-the-art detection systems, such as those by Rowley [43] and Viola and Jones [52]. Two possible causes have already been suggested, but there are additional factors that affect the template's performance. For instance, the template is sensitive to false positives because it uses a simple match threshold as its classifying function. A stronger match metric would greatly improve performance, perhaps to the point where it could be demonstrated competitively on public datasets such as the MIT-CMU image set. (The ratio-template was tested on the MIT-CMU images, but performance was relatively poor and it was decided not to include the results in the thesis.)

The second major limitation is that the detection scans are not as fast as they could be. Speed is directly proportional to the number of relations that are evaluated at each position in the image. Scanning a 340×240 pixel image without multiscale support and 90% detection accuracy took 0.53 seconds; with multiscale support, the scan took 2.6 seconds. The tradeoff between system speed and detection accuracy is a common but tenable problem. For example, the additional computational complexity of using a stronger, slower match metric could be offset by reducing the number of relations, many of which may be redundant due to excessive overlapping. (The problem of excessive overlap was briefly mentioned in section 3.1.3.)

Solutions to these limitations are addressed in a discussion on future work in section 4.3.

3.3 Summary

This chapter described a learning system that can automatically construct a ratio-template from a set of training images. The system semi-exhaustively extracts pairs of rectangular regions from the training images and evaluates pairs of regions as potential contrast relations using the relation rule. The relation rule specifies that region pairs must have above-threshold contrast whose polarity is invariant across all of the training examples.

Results were presented from a face detection task, in which the template's tolerance to noise and variation in illumination were demonstrated. It was shown that the template is capable of detecting faces at multiple scales. But the results also suggest that the ratio-template learning system must undergo significant optimizations to improve detection performance before it reasonably can be compared to existing high-performance face detection systems.

Chapter 4

Conclusion

This thesis has presented a sparse representation for image structure, called a ratio-template, that is quasi-invariant to changes in illumination and tolerant of image degradations such as sensor noise and resolution loss. Introduced in chapter 2, the ratio-template employs low-resolution ordinal contrast relationships that were inspired by our knowledge of biological vision systems. The relations encode the contrast polarity between pairs of spatially coarse image regions using a binary ordinal value (*is-darker-than* or *is-brighter-than*). Studies in neurophysiology provide support for this contrast encoding mechanism, showing that some cells in the primary visual cortex have large receptive fields and rapidly saturating contrast response functions that can be approximated by a step function. (The step function has only two states, hence the binary ordinal structure.) Perceptual psychology has shown the importance of contrast polarity in segmenting objects from their background and the sufficiency of low-resolution information in detecting objects such as faces.

Chapter 3 demonstrated that a ratio-template can be learned automatically from a set of examples. The template is constructed by extracting invariant contrast relations from a set of normalized images. One constructed, the template is exhaustively applied to a new image in search of instances of the learned object class. At each location in the image, the relations in the template are compared to the underlying image relations. The number of matching relations is thresholded to determine the presence or absence of an object.

Ordinal representations have been applied to many problems, including the detection of cars, pedestrians and faces, as well as image database indexing and retrieval. They are robust to noise and outliers because they do not depend on continuous quantitative values but rather on discrete, qualitative intervals. However, by discarding the absolute contrast values to achieve robustness, information is lost that was potentially useful in recognizing individuals within an object class.

Tests in chapter 3 on scanned faces demonstrated the ratio-template's detection accuracy and speed. The template can detect faces with 90% accuracy and fewer than one false positive in 1,000 images. But this is weaker than both the hand-crafted template and state-of-the-art face detection systems by Rowley [43] and Viola *et al.* [52]. In terms of speed, the unoptimized ratio template is a contender among other fast detectors. At 90% accuracy, it can process a 320×240 pixel image in 2.6 seconds at multiple scales. In comparison, Rowley's [43] upright face detector took one second to process a 320×240 image. Viola's [52] face detector took 0.067 seconds for a 384×288 image. These times are only representational: it is difficult to compare them directly, given the different hardware platforms, programming languages, code and compiler optimizations, etc., that were used in the various implementations. However, the ratio-template is relatively fast and potentially can be optimized to run faster without compromising its classification ability. Suggestions for improving both detection accuracy and speed are discussed below.

4.1 Strengths of the model

The ratio-template's strong points as an object representation include the following.

- The ratio-template sparsely encodes an object using a small collection of binary contrast relations. For example, in order to represent a 25×29 pixel image of a face, template RT-1.4 (chapter 3) required 127 relations or a total of 127 bits of information, making the template computationally inexpensive. In contrast, one of the example images consumes $25 \times 29 \times 8 = 5800$ bits.

- It is quasi-invariant to changes in illumination. The strength of the invariance depends on the diversity of illumination in the training examples. If the examples capture a wide range of illumination parameters, the resulting template will tolerate this range. Other methods typically approach illumination invariance as an afterthought—usually through normalization procedures—rather than as an integral part of the representation.

- Because the contrast relations coarsely segment the image feature space, training requires a relatively small number of examples, numbering in the hundreds, rather than in the thousands as is typical with many learning-based systems.

4.2 Weaknesses of the model

The ratio-template's weaknesses as an object representation are as follows.

- Due to its reliance on low-resolution features and ordinal metrics, the ratio-template is probably not suitable for identification tasks. This hypothesis has not been tested, but effective identification algorithms tend to rely on more descriptive, higher-dimensional representations [4, 12].

- Classification performance and speed are not yet high enough to be compared competitively to existing high-performance systems. Techniques for improving performance are discussed below.

- Although the ratio-template is a biologically plausible representation, there is currently no evidence that biological vision systems use such templates.

4.3 Future work

Two directions for future work on the ratio-template learning system have already been mentioned. First and foremost would be to improve the system's classification accuracy and test it on challenging public datasets such as the MIT-CMU image

collection. Suggestions made in section 3.2.5 included changing the training set and/or the normalization process. The possibility of using a different classifier was also mentioned. A stronger classifier could be used to analyze the pattern of matching image/template relations without changing the basic representation. Classifiers such as neural networks and support vector machines boast high accuracy but are poor choices because they require long training periods and thousands of training examples. They also add a significant layer of complexity which sharply contrasts with the simplicity of the underlying representation. A more appropriate choice might be a correlation function like that used by Bhat and Nayar [8]. Their function was derived expressly to evaluate matches between windows of ordinal values with high accuracy.

Another way to improve classification accuracy would be to augment the qualitative representation with some quantitative information. One possibility in this regard is to individually weight the contrast relations using their contrast magnitudes. This method might improve accuracy but it would also enlarge the search space, requiring more training examples to arrive at the proper weight values. Too heavy a reliance on quantitative values might also compromise the robustness of the representation.

The second direction for future work is optimizing the system's speed. The speed is proportional to the number of relations evaluated at each image position. Reducing this number would directly reduce the detection time. This could be accomplished by evaluating the relations in order of their importance. Relative importance might be based on the weighting scheme described above. Another way to reduce the number of evaluations would be to simply eliminate relations. The current implementation uses an overcomplete set of relations to improve accuracy, but some relations overlap so much that they are redundant. These, and possibly other, relations could be combined to reduce overcompleteness. (This is similar to the feature selection optimization used by Papageorgiou *et al.* [39].) Relations could also be eliminated by thresholding their weight values. Although these methods could potentially improve detection speed, any reduction in relations would need to be done carefully so as not to compromise detection performance.

Finally, in a test of the representation's ability to encode different classes of ob-

jects, it would be interesting to train ratio-templates on examples from different domains. The objects from these domains would need to bear characteristic signatures of amplitude variations, where amplitude may be defined over any of a variety of properties such as luminance, spatio-temporal frequency, color or entropy. Speech is one such domain, in which audio spectrograms could be analyzed for qualitative signatures correlated with phonemes or words.

Bibliography

[1] Edward H. Adelson, Charles H. Anderson, James R. Bergen, Peter J. Burt, and Joan M. Ogden. Pyramid methods in image processing. *RCA Engineer*, 29(6):33–41, November/December 1984.

[2] Duane G. Albrecht and David B. Hamilton. Striate cortex of the monkey and cat: Contrast response function. *Journal of Neuroscience*, 48(1):217–237, July 1982.

[3] Simon Baker, Shree K. Nayar, and Hiroshi Murase. Parametric feature detection. In *Proceedings of the 1997 DARPA Image Understanding Workshop*, pages 1425–1430, May 1997.

[4] Simon Baker, Shree K. Nayar, and Hiroshi Murase. Parametric feature detection. In *Proceedings of the 1997 DARPA Image Understanding Workshop*, pages 1425–1430, May 1997.

[5] Robert J. Baron. Mechanisms of human facial recognition. *International Journal of Man-Machine Studies*, 15(2):137–178, 1981.

[6] David J. Beymer. Face recognition under varying pose. Artificial Intelligence Laboratory Memo 1461, MIT, December 1993.

[7] David J. Beymer and Tomaso Poggio. Face recognition from one example view. Artificial Intelligence Laboratory Memo 1536, MIT, September 1995.

[8] Dinkar N. Bhat and Shree K. Nayar. Ordinal measures for image correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):415–423, April 1998.

[9] Martin Bichsel. *Strategies of Robust Object Recognition for Automatic Identification of Human Faces*. PhD thesis, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, 1991.

[10] Christopher M. Brislawn. Wavelet transforms and filter banks. `http://math.lanl.gov/ams/fy01/wavelets.html`.

[11] Vicki Bruce and Andy Young. *In the eye of the beholder: the science of face perception*. Oxford University Press, 1998.

[12] Roberto Brunelli and Tomaso Poggio. HyperBF networks for real object recognition. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1278–1285, Sydney, Australia, 1991.

[13] Roberto Brunelli and Tomaso Poggio. Face recognition: features versus templates. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15(10):1278–1284, 1993.

[14] Peter J. Burt. Multiresolution techniques for image representation, analysis, and 'smart' transmission. In *Proceedings of The International Society for Optical Engineering (SPIE) Volume 1199, Visual Communications and Image Processing IV*, pages 2–15, 1989.

[15] Marco La Cascia and Stan Sclaroff. Fast, reliable head tracking under varying illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 604–610, 1999.

[16] Patrick Cavanagh. What's up in top-down processing? In Andrei Gorea, editor, *Representations of Vision: Trends and Tacit Assumptions in Vision Research*, pages 295–304. Cambridge University Press, Cambridge, 1991.

[17] Peng Chang and John Krumm. Object recognition with color coocurrence histograms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 498–504, Fort Collins, Colorado, June 1999.

[18] William J. Conover. *Practical Nonparametric Statistics*. John Wiley, New York, 1980.

[19] Ian Craw, David Tock, and Alan Bennett. Finding face features. In *Proceedings of the Second European Conference on Computer Vision*, pages 92–96, Santa Margherita Ligure, Italy, May 1992.

[20] Gregory C. DeAngelis, Izumi Ohzawa, and Ralph D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *Journal of Neurophysiology*, 69(4):1091–1117, April 1993.

[21] Mark S. Drew, Jie Wei, and Ze-Nian Li. Illumination invariant color object recognition via compressed chromaticity histograms of color-channel-normalized images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 533–540, Bombay, January 1998.

[22] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience, New York, 1973.

[23] Brian V. Funt and Graham D. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5), May 1995.

[24] Ruth Ellen Galper and Julian Hochberg. Repetition memory for photographs. *American Journal of Psychology*, 84(3):351–354, 1971.

[25] Theo Gevers. Robust histogram construction from color invariants. In *Proceedings of the Eighth IEEE International Conference on Computer Vision*, volume 1, pages 615–620, Vancouver, July 2001.

[26] Venu Govindaraju. Locating human faces in photographs. *International Journal of Computer Vision*, 19(2):129–146, 1996.

[27] Amara Graps. An introduction to wavelets. *IEEE Computational Sciences and Engineering*, 2(5):50–61, 1995.

[28] W. Eric L. Grimson and Tomas Lozano-Perez. Localizing overlapping parts by searching the interpretation tree. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 9(4):469–482, 1987.

[29] Peter Halinan, Alan Yuille, and David Mumford. Harvard Face Database.

[30] Leon D. Harmon. The recognition of faces. *Scientific American*, 227:71–82, November 1973.

[31] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–768, Puerto Rico, June 1997.

[32] Daniel Huttenlocher, Gregory Klanderman, and William Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Patter Analysis and Machine Intelligence*, 15(9):850–863, September 1993.

[33] Qasim Iqbal and J. K. Aggarwal. Combining structure, color and texture for image retrieval: A performance evaluation. In *To appear in the IEEE International Conference on Pattern Recognition*, Quebec City, Canada, August 2002.

[34] Michael Kass, Andy Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[35] Tai Sing Lee. Image representation using 2D Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971, 1996.

[36] Pamela R. Lipson. *Context and Configuration Based Scene Classification*. PhD thesis, Massachusetts Institute of Technology, 1996.

[37] Tom M. Mitchell. *Machine Learning*, pages 67, 111. MIT Press and WCB McGraw-Hill, 1997.

[38] David W. Murray and D.B. Cook. Using the orientation of fragmentary 3D edge segments for polyhedral object recognition. *International Journal of Computer Vision*, 2(2):153–169, 1988.

[39] Constantine P. Papageorgiou and Tomaso Poggio. Trainable pedestrian detection. In *Proceedings of the International Conference on Image Processing*, pages 25–28, Kobe, Japan, October 1999.

[40] Richard J. Phillips. Why are faces hard to recognize in photographic negative? *Perception and Psychophysics*, 12(5):425–426, 1972.

[41] B. Lucier R. DeVore, B. Jawerth. Image compression through wavelet transform coding. *IEEE Transactions on Information Theory*, 38(2):719–746, 1992.

[42] Aparna Lakshmi Ratan. *Learning Visual Concepts for Image Classification*. PhD thesis, Massachusetts Institute of Technology, 1999.

[43] Henry A. Rowley. *Neural Network-based Face Detection*. PhD thesis, Carnegie Mellon University, May 1999.

[44] Karl Schwerdt and James L. Crowley. Robust face tracking using color. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 90–95, Grenoble, France, March 2000.

[45] Pawan Sinha. Qualitative image-based representations for object recognition. Artificial Intelligence Laboratory Memo 1505, MIT, 1994.

[46] Eric J. Stollnitz, Tony D. DeRose, and David H. Salesin. Wavelets for computer graphics: A primer. Technical Report 94-09-11, University of Washington, Department of Computer Science and Engineering, September 1994.

[47] Kah-kay Sung and Tomaso Poggio. Example-based learning for view-based human face detection. Artificial Intelligence Laboratory Memo 1521, MIT, December 1994.

[48] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[49] Simon Thorpe. `http://www.spikenet-technology.com/English/concept.htm`. SpikeNet, Inc., Toulouse, France, 1999.

[50] Antonio Torralba and Pawan Sinha. Detecting faces in impoverished images. Artificial Intelligence Laboratory Memo 2001-028, MIT, November 2001.

[51] Alessandro Verri, Marco Straforini, and Vincent Torre. Computational aspects of motion perception in natural and artificial vision systems. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 337(1282):429–443, 1992.

[52] Paul Viola and Michael Jones. Robust real-time object detection. In *Proceedings of the Second International Workshop on Statistical and Computational Theories of Vision*, Vancouver, July 2001.

[53] Roger Watt. A computational examination of image segmentation and the initial stages of human vision. *Perception*, 23:383–398, 1994.

[54] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proceedings of the European Conference on Computer Vision*, pages 151–158, 1994.